SEQUENTIAL METHODS FOR LEARNING AND INFERENCE UNDER UNKNOWN MODELS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Sattar Vakili

August 2017

SEQUENTIAL METHODS FOR LEARNING AND INFERENCE UNDER UNKNOWN MODELS

Sattar Vakili, Ph.D.

Cornell University 2017

This dissertation focuses on sequential learning and inference under unknown models. In this class of problems, observations are available one at a time, and optimal actions cannot be determined *a priori*. Solutions to such problems require optimization over time: to learn optimally from the past and then act accordingly with foresighted planning.

The first part of the dissertation is on sequential learning within the framework of multi-armed bandit (MAB) theory. Originated in 1930s, the MAB problems have been traditionally studied under the following modeling assumptions: the reward distributions are assumed to have a bounded support, the learning is centralized with a single player, the reward distributions are assumed to be time-invariant, and the performance measure mainly targets at maximizing the expected return of a sequential learning policy. This dissertation aims to relax these modeling assumptions and develop parallel results applicable to a wider range of applications, in particular, emerging applications such as online ad placement and web search as well as social and economic networks.

The main results in the first part include a new policy based on a deterministic sequencing of exploration and exploitation for the classic MAB. The structure of the proposed policy allows easy extensions to variations of MAB, including decentralized MAB with multiple players and incomplete reward observa-

tions under collisions. The proposed policy achieves the optimal logarithmic *regret* order under general reward distribution models such as Sub-Gaussian and heavy-tailed distributions. The time-variation in the reward process is also addressed by considering an arbitrary time-variation as well as a piece-wise stationary model. A lower bound on the regret order is obtained and order optimality of a sequential learning policy is shown. The issue of risk in multi-armed bandit problems is considered and parallel results are developed under the measure of mean-variance, a commonly adopted risk measure in economics and mathematical finance. It is shown that the model-specific regret and the model-independent regret in terms of the mean-variance of the reward process are lower bounded by $\Omega(\log T)$ and $\Omega(T^{\frac{2}{3}})$, respectively. It is also shown that variations of the policies developed for the classic risk-neutral MAB achieve these lower bounds. Also, under the measure of value at risk (another common risk measure in financial mathematics) a sequential learning policy is developed. In addition, a minimal side information on the reward model is introduced that can lead to bounded regret, thus, complete learning. Provided the side information, a sequential learning policy with bounded regret is proposed.

The second part of the dissertation is on sequential inference which can be categorized into three classes of problems: sequential hypothesis testing, change detection, and active hypothesis testing. These classic problems date back to 1940s with pioneering works by Wald, Chernoff, Page, and Shewhart. The focus of this dissertation is on time-varying models, hierarchical observations, and non-parametric composite hypotheses motivated by recent engineering applications such as short-term instability detection from PMU measurements in power systems and detection of heavy hitters and denial-of-service attacks in the Internet, and communication and computer networks.

The main results in the second part include asymptotically optimal tests for the sequential hypothesis testing and change-point detection problems under a time-varying distribution model suitable for instability detection applications. The asymptotic optimality of the tests are proven by establishing fundamental limits on the Bayesian cost of any test. The problem of active inference under hierarchical observations, and unknown and non-parametric models is also considered. A policy that interactively chooses the observation points is proposed and is shown to achieve optimal logarithmic order sample complexity in both the problem size and the reliability constraint.

To my family

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to all who have inspired and challenged me during my graduate studies. This achievement could not happen without care, love, encouragement and constructive criticism I have received from my professors, colleagues and friends.

First and foremost, I am very grateful to my advisor professor Qing Zhao for her continuous guidance and support throughout many years. Her professional manner, great insight, enthusiasm for learning and teaching, incredible patience, and outstanding advisory have contributed a major share to both my professional and personal life as a graduate student. I have enjoyed and acquired so many academic skills in our weekly meetings. I cannot thank her enough.

I would also like to extend my gratitude to professors Aaron Wagner, Eilyan Bitar, and Lang Tong for their incredible support and valuable discussions and suggestions during my time at Cornell.

I would like to thank my colleagues who have always been a great inspiration for me. I would like to thank Carlos, Lydia, Arash, Chao, Sogol, Jianhang, Saghar, Nariman, Kia, Marie, Najva and many other friends and colleagues for making my time as a graduate student fun and memorable.

I am so grateful to my family; to my parents for their unconditional love and support, and to my siblings for their charm, wit and sense of humor.

Last but not least, I am so grateful to Hanifa whose love, intellect, and authenticity have made the last year of the graduate school for me the most lively and colorful.

TABLE OF CONTENTS

Dedication 6
Acknowledgements 7
Table of Contents 8
List of Figures 12

1 Introduction **1**
1.1 Sequential Learning 1
1.2 Sequential Inference 4
1.3 Organization of the Dissertation 5

I Sequential Learning **7**

2 Literature Review on Multi-Armed Bandit Problems **8**
2.1 Problem Formulation 9
2.1.1 Model Specific and Model Independent Regret 10
2.2 Classic Results 11
2.2.1 Lower Bounds 11
2.2.2 Policies 12
2.3 Variations of MAB 14

3 A New Policy for Various Multi-Armed Bandit Problems under General Distribution Models **16**
3.1 The DSEE Policy 19
3.1.1 The General Structure 19
3.1.2 Preliminaries 21
3.1.3 Under Sub-Gaussian Reward Distributions 22
3.1.4 Sublinear Regret under Heavy-Tailed Distribution Model with Sublinear Complexity and No Prior Knowledge . . . 26
3.1.5 Logarithmic Regret under Heavy-Tailed Distribution Model using Truncated Sample Mean 28
3.2 Extendibility to Variations of MAB 31
3.2.1 MAB under Various Objectives 31
3.2.2 Decentralized MAB with Incomplete Reward Observations 33
3.3 Conclusion 36

4 Time-Varying Stochastic Multi-Armed Bandit Problems **38**
4.1 Problem Formulation and Preliminaries 41
4.2 Weak Regret 42
4.3 Piece-wise Stationary Time Variation Model 44
4.4 Conclusion 46

5 Risk-Averse Multi-Armed Bandit under Mean-Variance Measure 47
5.1 Motivating Applications . . . 50
5.2 Related Work . . . 51
5.3 Notation and Preliminaries . . . 53
5.3.1 Notations . . . 53
5.3.2 Concentration of the Sample Mean-Variance . . . 55
5.4 The Known Model Case . . . 56
5.5 Model-Specific Regret . . . 59
5.5.1 Lower Bounds on Model-Specific Regret . . . 59
5.5.2 Risk-Averse Learning Policies . . . 62
5.6 Model-Independent Regret . . . 64
5.7 Simulations . . . 66
5.8 Discussion . . . 66

6 Risk-Averse Multi-Armed Bandit under Value at Risk Measure 70
6.1 Notation and Preliminaries . . . 72
6.2 Gaussian Reward Distribution Model . . . 73
6.3 Sub-Gaussian Reward Distribution Model . . . 77
6.4 Conclusion . . . 80

7 Achieving Bounded Regret with Minimal Side Information 81
7.1 Bounded Regret Policy . . . 82
7.1.1 Under Sub-Gaussian Distribution Model . . . 82
7.1.2 Under Heavy-Tailed Distribution Model . . . 84
7.2 Conclusion . . . 88

II Sequential Inference 89

8 Literature Review on Sequential Inference 90
8.1 Sequential Hypothesis Testing . . . 90
8.2 Quickest Change-Point Detection . . . 93
8.3 Active Inference . . . 95

9 Sequential Hypothesis Testing and Change Detection under Time-Varying Models 96
9.1 Motivation: Voltage Instability Detection in Power Systems . . . 98
9.2 Binary Hypothesis Testing: Problem Formulation . . . 100
9.3 Binary Hypothesis Tests . . . 101
9.3.1 Under Simple Hypothesis Model . . . 101
9.3.2 Under Composite Hypothesis Model . . . 103
9.3.3 Simulations . . . 106
9.4 Bayesian Quickest Change-point Detection: Problem Formulation 107
9.5 Quickest Change-Point Detection Tests . . . 108

9.5.1 Under Known Post-Change Parameter . . . 108
9.5.2 Under Unknown Post-Change Parameter . . . 111
9.5.3 Simulations . . . 113
9.6 Instability Detection in General Linear Systems . . . 113
9.7 Conclusion . . . 116

10 Active Inference under Hierarchical Observations and Unknown Models 117
10.0.1 Applications . . . 118
10.0.2 Main Results . . . 120
10.0.3 Related Work . . . 122
10.1 Problem Formulation . . . 124
10.2 An Active Inference Strategy: CBRW . . . 127
10.2.1 Detecting Leaf-Level Targets . . . 128
10.2.2 Detecting Hierarchical Targets . . . 130
10.3 Performance Analysis . . . 131
10.3.1 Leaf-Level Target Setting . . . 132
10.3.2 Hierarchical Target Setting . . . 134
10.4 Extensions . . . 134
10.4.1 Multiple Targets . . . 135
10.4.2 Heavy-Tailed Distributions . . . 136
10.4.3 General Tree Structure . . . 137
10.4.4 Variations for Different Applications . . . 137
10.5 Conclusion . . . 140

A Proofs of Lemmas and Theorems from Part I 142
A.1 Proof of Lemma 2 . . . 142
A.2 Proof of Lemma 3 . . . 143
A.3 Proof of Lemma 4 . . . 147
A.4 Proof of Theorem 12 . . . 148
A.5 Proof of Lemma 5 . . . 150
A.6 Proof of Theorem 13 . . . 153
A.7 Proof of Lemma 6 . . . 157
A.8 Proof of Theorem 14 . . . 158
A.9 Proof of Theorem 16 . . . 158
A.10 Proof of Theorem 8 . . . 162
A.11 Proof of Theorem 23 . . . 165
A.12 Proof of Theorem 10 . . . 166
A.13 Proof of Theorem 11 . . . 167

B Proofs of Lemmas and Theorems from Part II 169
B.1 Proof of Theorem 25 . . . 169
B.2 Proof of Theorem 26 . . . 173
B.3 Proof of Theorem 27 . . . 176

B.4 Proof of Lemma 9 182
B.5 Proof of Theorems 28 and 29 184

Bibliography **189**

LIST OF FIGURES

3.1 The DSEE approach for the classic MAB. 20
3.2 An example of decentralized policies based on DSEE ($M = 2, K = 3$, the index of the selected arm at each time is given). 36

4.1 The EXP3 Algorithm 42
4.2 The EXP3.S Algorithm 44

5.1 The sample observations of MV-UCB under different risk-tolerance factor ρ. 66
5.2 The performance of MV-UCB ($\rho = 1$, $K = 2$ with normal reward distributions of parameters $\mu_1 = 0, \mu_2 = 0.5, \sigma_2^2 = 1$). 67

7.1 The description of the π_η policy. 83

9.1 The probability of error for SGLRT-exp. 105
9.2 The average sample number. 105
9.3 The probability of error for SGLRT-exp. 106
9.4 The average sample number. 107
9.5 Probability of error. $\rho = 0.01$. 114
9.6 Probability of error. $\rho = 0.1$. 114
9.7 Expected voltage instability detection delay. $\rho = 0.01$. 115
9.8 Expected voltage instability detection delay. $\rho = 0.1$. 115

10.1 The hierarchical data streams model. 118
10.2 The hierarchical data streams model. 133
10.3 Finite-regime upper bounds on the performance of CBRW under leaf-level 10.17 and hierarchical 10.18 target settings. 135
10.4 An example of a general (not necessarily binary) hierarchical data streams model. 138

CHAPTER 1

INTRODUCTION

This dissertation focuses on sequential learning and inference under unknown models. In this class of problems, observations are available one at a time, and optimal actions cannot be determined *a priori*. Solutions to such problems require optimization over time: to learn optimally from the past (previous observations) and then act accordingly with foresighted planning.

The results are presented in two parts. The first part is on sequential learning problems. The second part is on sequential inference problems. In sequential inference problems the focus is on the accuracy of a terminal inference. While, in sequential learning problems, the cumulative value of a sequential reward (or cost) is optimized.

1.1 Sequential Learning

The first part of the results presented in this dissertation is on sequential learning problems within the framework of Multi-Armed Bandit (MAB). Originated in 1930s, the MAB is a class of sequential learning and decision-making problems under unknown models. An abstraction of this class of problems involves a slot machine with K independent arms and a single player. At each time, the player chooses one arm to play and obtains a random reward drawn i.i.d. over time from an unknown distribution specific to the chosen arm. The design objective is a sequential arm selection policy that maximizes the total expected reward over a horizon of length T by striking a balance between learning the unknown reward models of all arms (exploration) and capitalizing on this

information to earn immediate reward (exploitation).

The performance of an arm selection policy is measured by *regret* defined as the expected cumulative reward loss over the entire time horizon against an omniscient player who knows the reward models and always plays the best arm [1]. In their seminal work [1], Lai and Robbins showed that the minimum regret growth rate is $\Omega(\log T)$. Several online learning policies exist in the literature that achieve the optimal regret order under various assumptions on the reward models (see [1, 2, 3, 4, 5]).

Our contribution to this literature consists of introducing new settings, proposing novel policies and providing analysis for several extensions of the classic formulation which allows for a wider applicability of the results.

We first propose a new policy referred to as Deterministic Sequencing of Exploration and Exploitation (DSEE) for the classic MAB. The deterministic structure of the proposed policy allows easy extensions to variations of MAB, including decentralized MAB with multiple players and incomplete reward observations under collisions. We provide a comprehensive analysis of the proposed policy under several distribution models including heavy-tailed distributions. We show that DSEE achieves optimal logarithmic order regret under a Sub-Gaussian distribution model. For heavy-tailed reward distributions, we show that DSEE achieves a sublinear regret. We also show that with the knowledge of an upper bound on a finite moment of the heavy-tailed reward distributions, DSEE offers the optimal logarithmic regret order.

We then address the time variation in the reward process. In particular, we consider a time-varying stochastic multi-armed bandit (MAB) problem with ar-

bitrary unknown distribution models assigned by an adversary. We obtain a lower bound on the regret order and show that an online learning algorithm achieves this lower bound. We further consider a piece-wise stationary model for the reward distributions and analyze the regret performance of an online learning policy in terms of the number of change points experienced by the reward distributions over the time horizon.

Under a model with a different performance measure, we study risk-averse MAB problems. The classic MAB formulation mainly targets at maximizing the expected return of an online learning policy. In many applications, especially in economics and finance, a decision maker may be more interested in reducing the uncertainty (i.e., risk) in the outcome, rather than achieving the highest ensemble average. We develop novel results parallel to those on classic MAB under the measure of mean-variance, a commonly adopted risk measure in economics and mathematical finance. We show that the model-specific regret and the model-independent regret in terms of the mean-variance of the reward process are lower bounded by $\Omega(\log T)$ and $\Omega(T^{2/3})$, respectively. We then show that variations of the UCB policy and the DSEE policy developed for the classic risk-neutral MAB achieve these lower bounds. We also consider MAB under the measure of Value at Risk and develop a learning policy under certain assumptions.

In addition, we introduce a minimal side information that makes the complete learning possible. The logarithmic regret growth rate shown by Lai and Robbins indicates that a learning policy never achieves complete learning and inevitably keeps choosing suboptimal arms at a certain rate with time. We demonstrate the possibility of achieving complete learning provided a minimal

side information.

1.2 Sequential Inference

The second part of the results presented in this work is on sequential inference problems. Provided sequences of observations from an environment, the objective is to detect particular underlying phenomenons with the smallest possible number of observations. The essence of the problems is the tension between the delay and the reliability: the desired reliability can be achieved through the accumulation of measurements, which comes at the price of increasing the detection delay.

The sequential inference problems considered here can be categorized into three classes of problems: sequential hypothesis testing, change detection, and active hypothesis testing. The classic sequential hypothesis testing problem was pioneered by Wald [6]. Wald showed that the sequential probability ratio test (SPRT) is optimal in terms of minimizing the expected sample size subject to given error probability constraints. Shiryaev in 1960's [7, 8] developed an optimal Bayesian change-point detection policy. The problem of sequential design of experiments where a set of different experiments are available and the observations depend on the chosen experiment was studied by Chernoff [9].

The focus of this dissertation is on time-varying models, hierarchical observations, and non-parametric composite hypotheses motivated by recent engineering applications such as short-term instability detection from PMU measurements in power systems and detection of heavy hitters and denial-of-service attacks in the Internet, and communication and computer networks.

We address the sequential hypothesis testing and change-point detection problems under a time-varying distribution model suitable for instability detection applications. In this model, the mean value of the observation process has an exponential dependence on time. We propose sequential tests and prove their asymptotic optimality by establishing fundamental limitations for any sequential test.

We also consider the problem of active inference under hierarchical observations and unknown and non-parametric models. The decision maker is allowed to interactively choose the observation points. We show a logarithmic sample complexity in problem size (provided hierarchical observations) as well as in reliability constraint. The general active inference problem studied in this work finds variety of applications and closely relates to several other problems including group testing, heavy hitter detection in Internet monitoring, and active learning with binary threshold classifiers.

1.3 Organization of the Dissertation

The rest of the dissertation is presented in two parts. Part I is dedicated to the results on online learning problems. This part consists of Chapters 2, 3, 4, 5, 6 and 7. In Chapter 2, we introduce the MAB problem and review the existing results. In Chapter 3, we introduce the DSEE policy. We provide analysis for the regret performance of DSEE under both Sub-Gaussian and heavy-tailed reward distributions and discuss the extendability of the results to several variations of the classic MAB. In Chapter 4, we consider a MAB problem with time-varying distribution model. We consider a setting where the distribution

model is allowed to change arbitrarily and analyze the weak regret under this setting. We also consider a different model with piece-wise stationary distributions. In Chapter 5, we present the risk-averse MAB problem under the measure of mean-variance. We establish novel results on the lower bounds for the risk-averse regret. We also analyze the performance of two risk-averse policies and show their order optimality. In Chapter 6, we develop a learning policy for a risk-averse MAB problem under the measure of value at risk. In Chapter 7, we show that provided a value between the mean values of the optimal and next to optimal arm, bounded regret (thus, complete learning) is achievable.

Part II is dedicated to the results on sequential inference. This part consists of Chapters 8, 9, and 10. In Chapter 8, we review the existing results on sequential inference. In Chapter 9, we study sequential hypothesis testing and quickest change-point detection problems under a non-stationary distribution model. We demonstrate the applicability of the results to instability detection in linear and some non-linear systems (power systems in particular). In Chapter 10, we introduce a general active inference problem under hierarchical observations and unknown models. We propose a sequential target search policy that performs as a random walk and provide the performance analysis. We discuss several other sequential decision problems which closely relate to our problem.

The proof of lemmas and mathematically more involved theorems are provided in the Appendices. Appendix A and Appendix B, respectively, include the proofs of theorems and lemmas from Part I and Part II which are not presented in the main text.

Part I

Sequential Learning

CHAPTER 2

LITERATURE REVIEW ON MULTI-ARMED BANDIT PROBLEMS

Multi-armed bandit (MAB) is a class of online learning and sequential decision-making problems under unknown models. An abstraction of this class of problems involves a slot machine with K independent arms and a single player. At each time, the player chooses one arm to play and obtains a random reward drawn i.i.d. over time from an unknown distribution specific to the chosen arm. The design objective is a sequential arm selection policy that maximizes the total expected reward over a horizon of length T by striking a balance between earning immediate reward (exploitation) and learning the unknown reward models of all arms (exploration).

In the MAB problem, each received reward plays two roles: increasing the wealth of the player, and providing one more observation for learning the reward statistics of the arm. The tradeoff between exploration and exploitation is thus clear: which role should be emphasized in arm selection—an arm less explored thus holding potentials for the future or an arm with a good history of rewards?

The MAB problem dates back to thirties when it was proposed for clinical trial applications. In 1952, Robbins addressed the two-armed bandit problem [10]. He showed that the same maximum average reward achievable under a known model can be obtained by dedicating two arbitrary sublinear sequences for playing each of the two arms. In 1985, Lai and Robbins proposed a finer performance measure, the so-called regret, defined as the expected total reward loss with respect to the ideal scenario of known reward models (under which the best arm is always played) [1]. Regret not only indicates whether the maximum average

reward under known models is achieved, but also measures the convergence rate of the average reward, or the effectiveness of learning. Although all policies with sublinear regret achieve the maximum average reward, the difference in their total expected reward can be arbitrarily large as T increases. The minimization of the regret is thus of great interest. Lai and Robbins showed that the minimum regret has a logarithmic order in T.

2.1 Problem Formulation

Consider a K-armed bandit and a single player. At each time t, the player chooses one arm to play. Playing arm k yields a random reward $X_k(t)$ drawn i.i.d. from an unknown distribution f_k. Let $\mathcal{F} = (f_1, \cdots, f_K)$ denote the set of the unknown distributions. An arm selection policy π specifies a function at each time t that maps from the player's observation and decision history to the arm to play at time t. Let $\{X_{\pi(t)}(t)\}_{t=1}^{T}$ denote the random reward sequence under policy π.

The performance of policy π is measured by regret $R_\pi(T)$ defined as the expected cumulative reward loss over the entire time horizon against an omniscient player who knows the reward models and always plays the best arm:

$$R_\pi(T) = \mathbb{E}_{\mathcal{F}}[\sum_{t=1}^{T} X_*(t) - \sum_{t=1}^{T} X_{\pi(t)}(t)]. \tag{2.1}$$

Where $*$ is used to denote the optimal arm with the largest mean value: $\mu_* = \max_k \mu_k$. It is assumed that there is a single optimal arm $*$; otherwise one of them will be chosen arbitrarily. The notation $\mathbb{E}_{\mathcal{F}}$ denotes the expectation operator with respect to the distribution model $\mathcal{F}$. Let $\tau_k(t)$ denote the number of times

arm k has been played up to time t. In order to fully specify $\tau(t)$, indeed, the arm selection policy needs to be specified. However, for simplicity of notations we drop the indication of policy from the notation of τ and specify it in the text if it is not clear. Let $\Delta_k = \mu_* - \mu_k$ be the gap in the mean values of the best arm and arm k. Based on Wald identity, the regret can also be written as

$$R_\pi(T) = \sum_{k=1}^{K} \mathbb{E}_\mathcal{F} \tau_k(T) \Delta_k. \tag{2.2}$$

The objective is to design arm selection policies π to minimize the regret. The policies are often designed to minimize the regret growth rate with time t. We often refer to a policy achieving logarithmic order regret as an order-optimal policy.

2.1.1 Model Specific and Model Independent Regret

The regret performance of a policy in the stochastic MAB can be evaluated under two settings: model specific and model-independent regret. The classic result by Lai and Ronbbins is under the former. In this setting, the regret performance is characterized specific to the given set of reward distributions. In establishing the lower bound for the model specific setting, there is an issue of trivial lower bounds on regret caused by policies that heavily bias toward specific arms. For example, a policy that always plays arm 1 has a 0 regret under a distribution model where arm 1 is the best arm. Lai and Robbins' result avoids such trivial lower bounds for model-specific regret by specifying the set of policies to be the so called uniformly good policies. A uniformly good policy is a

policy that archives $o(T^\alpha)$ regret for all $\alpha > 0$ under all legitimate distribution assignments. The Lai and Robbins' lower bound on regret is satisfied for any specific one-parameter distribution model and all uniformly good policies.

Subsequent studies on MAB considered also the model-independent setting in which the performance of a learning policy is measured against the worst-case assignment of the reward distributions. This setting does not face the issue mentioned above and the results generally hold for all arm selection policies.

2.2 Classic Results

The model-specific and model-independent lower bounds on regret for the classic MAB are in $\Theta(\log T)$ and $\Theta(\sqrt{T})$, respectively. Online learning policies have also been developed that obtain order optimal regret.

2.2.1 Lower Bounds

Lai and Robbins specifically showed that under a one parameter distribution model the regret of any uniformly good policy satisfies

$$\liminf_{T\to\infty} \frac{R_\pi(T)}{\log T} \geq \sum_{k\neq *} \frac{\Delta_k}{I_{k,*}}, \tag{2.3}$$

where Δ_k, the gap in the mean values of arm k and $*$, indicates the suboptimality of arm k, and $I_{k,*}$ denotes the Kullback-Leibler (KL) divergence between the distributions of k and $*$. In Lai and Robbis' lower bound, the number of times

a suboptimal arm k is played is proportional to $\frac{1}{I_{k,*}}$. This shows closer the distributions are (in terms of KL divergence) the more difficult is to distinguish them. Lai and Robbins also constructed explicit policies to achieve the minimum regret growth rate for several reward distributions including Bernoulli, Poisson, Gaussian, and Laplace [1] under the assumption that the distribution type is known. In [11], Bubeck *et al.* proved logarithmic lower bounds for the model-specific regret assuming the mean value of the best arm (μ_*) or the gap in the mean values of the best and the second best arm in known.

Under the worst case assignment of the distributions, the order of the lower bound in T is different than logarithmic. From the lower bound results in [11] and also from the lower bound results on the non-stochastic MAB problem studied in [12], an $\Omega(\sqrt{KT})$ lower bound on regret can be concluded. Specifically, for particular distribution assignments, a lower bound on the expected value of the number $\tau_k(T)$ of times a suboptimal arm k is played can be shown as:

$$\mathbb{E}\tau_k(T) \geq \frac{c}{\Delta_k^2} \log T \tag{2.4}$$

where c is a positive constant. The key idea is to consider two different distribution models where the distribution of the best arm and arm k are switched under these two models while the distribution of other arms remain the same. The worst case assignment of $\Delta_k = \sqrt{\frac{K}{T}}$ results in $\Omega(\sqrt{KT})$ lower bound for model-independent regret.

2.2.2 Policies

In [2], Agrawal developed index-type policies in explicit form for the distribution types considered in [10] as well as exponential distribution assuming

known distribution type. In [3], Auer *et al.* developed order optimal index policies for any unknown distribution with bounded support assuming the support range is known.

In these classic policies, arms are prioritized according to two statistics: the sample mean $\bar{\mu}(t)$ calculated from past observations up to time t and the number $\tau(t)$ of times that the arm has been played up to t. The larger $\bar{\mu}(t)$ is or the smaller $\tau(t)$ is, the higher the priority given to this arm in arm selection. The tradeoff between exploration and exploitation is reflected in how these two statistics are combined together for arm selection at each given time t. This is most clearly seen in the UCB (that stands for Upper Confidence Bound) policy proposed by Auer *et al.* in [3], in which an index $I(t)$ is computed for each arm and the arm with the largest index is chosen. The index has the following simple form:

$$I(t) = \bar{\mu}(t) + \sqrt{2\frac{\log t}{\tau(t)}}. \tag{2.5}$$

This index form is intuitive in the light of Lai and Robbins's result on the logarithmic order of the minimum regret which indicates that each arm needs to be explored on the order of $\log t$ times. For an arm sampled at a smaller order than $\log t$ times, its index, dominated by the second term, will be sufficient large for large t to ensure further exploration. The index $I(t)$ is an upper confidence bound for the actual mean value. Auer *et al.* [3] showed that for a bounded support distribution model (where the support of all distributions is limited to $[0, 1]$ interval) the UCB policy achieve a regret $R(T)$ no greater than

$$8\sum_{k\neq *}\frac{\log T}{\Delta_k} + (1+\frac{\pi^2}{3})\sum_k \Delta_k, \tag{2.6}$$

where $\Delta_* = \mu_* - \mu_k$ is the gap in the mean value of the optimal arm and arm k.

The model-specific regret bound obtained by UCB policy is in the optimal logarithmic order. However, the constants in front of log differ from the ones

provided by Lai and Robbins' lower bound. Later, Garivier and Cappe [13] proposed the so-called KL-UCB policy and proved better constants. Specifically, they showed that KL-UCB obtains a regret no bigger than

$$\sum_{k \neq *} \frac{\alpha \Delta_k}{I_{k,*}} \log T + O(1). \tag{2.7}$$

where $\alpha > 1$ is a parameter of the KL-UCB policy.

In [61, 62] the UCB policy was extended to all light-tailed distribution models. In [24], a modified version of UCB was shown to achieve the optimal logarithmic order regret under a heavy-tailed distribution model assuming the knowledge of an upper bound on a $p > 1$ moment of the distributions.

The above policies show an $O(\sqrt{KT} \log T)$ regret order under the worst case distribution assignment. There is thus a $\log T$ gap in the performance of the policies and the lower bound under model independent regret setting. A modification of the UCB policy, referred to as Improved UCB, was introduced by Auer and Ortner in 2010 and was shown to achieve the optimal model-independent regret order [5].

2.3 Variations of MAB

One variation of the classic formulation is decentralized MAB with multiple players. The problem was considered in [14] with a simple collision model: regardless of the occurrence of collisions, each player always observes the actual reward offered by the selected arm. In this case, collisions affect only the immediate reward but not the learning ability. It was shown that the optimal system regret has the same logarithmic order as in the classic MAB with a sin-

gle player, and a Time-Division Fair sharing (TDFS) framework for constructing order-optimal decentralized policies was proposed. Under the same complete observation model, decentralized MAB was also addressed in [15, 16], where the UCB was extended to the multi-player setting under a Bernoulli reward model. In [17], Tekin and Liu addressed decentralized learning under general interference functions and light-tailed reward models. In [18, 19], Kalathil *et al.* considered a more challenging case where arm ranks may be different across players and addressed both i.i.d. and Markov reward models. They proposed a decentralized policy that achieves near-$O(\log^2 T)$ regret for distributions with bounded support.

Bubeck *et al.* provided a comprehensive survey on different MAB settings and policies [20]. There is a large body of work on stochastic MAB problems under different variations and for various applications, including clinical trials, internet advertising, web search, and communication networks (see [21, 22, 23, 24, 25, 26, 27] and references therein). MAB has also been applied to a variety of scenarios in finance and economics (see, for example, a comprehensive survey in [28]).

CHAPTER 3

A NEW POLICY FOR VARIOUS MULTI-ARMED BANDIT PROBLEMS UNDER GENERAL DISTRIBUTION MODELS

In this chapter, we present a new approach to the MAB problem based on a Deterministic Sequencing of Exploration and Exploitation (DSEE). The DSEE approach differs from the classic policies proposed in [1, 2, 3] by separating in time the two objectives of exploration and exploitation. Specifically, time is divided into two interleaving sequences: an exploration sequence and an exploitation sequence. In the former, the player plays all arms in a round-robin fashion. In the latter, the player plays the arm with the largest sample mean (or a properly chosen mean estimator). Under this approach, the tradeoff between exploration and exploitation is reflected in the cardinality of the exploration sequence. It is not difficult to see that the regret order is lower bounded by the cardinality of the exploration sequence since a fixed fraction of the exploration sequence is spent on suboptimal arms. Nevertheless, the exploration sequence needs to be chosen sufficiently dense to ensure effective learning of the best arm. The key issue here is to find the minimum cardinality of the exploration sequence that ensures a reward loss in the exploitation sequence caused by incorrectly identified arm rank having an order no larger than the cardinality of the exploration sequence.

We show that when the reward distributions are Sub-Gaussian, DSEE achieves the optimal logarithmic order of the regret using an exploration sequence with $O(\log T)$ cardinality. For heavy-tailed reward distributions, DSEE achieves $O(T^{1/p})$ regret when the moments of the reward distributions exist up to the pth order for $1 < p \leq 2$ and $O(T^{1/(1+p/2)})$ for $p > 2$. More importantly,

with the knowledge of an upper bound on a finite moment of the heavy-tailed reward distributions and using proper mean estimators (to be specified later), DSEE offers the optimal logarithmic regret order.

We point out that both the classic policies in [1, 2, 3] and the DSEE approach developed in this work require certain knowledge on the reward distributions for policy construction. The classic policies in [1, 2, 3] apply to specific distributions with either known distribution types [1, 2] or known finite support range [3]. The advantage of the DSEE approach is that it applies to any distribution without knowing the distribution type. The caveat is that it requires the knowledge of a positive lower bound on the difference in the reward means of the best and the second best arms. This can be a more demanding requirement than the distribution type or the support range of the reward distributions. By increasing the cardinality of the exploration sequence, however, we show that DSEE achieves a regret arbitrarily close to the logarithmic order without *any* knowledge of the reward model. We further emphasize that the sublinear regret for reward distributions with heavy tails is achieved without any knowledge of the reward model (other than a lower bound on the order of the highest finite moment).

Different from the classic policies proposed in [1, 2, 3], the DSEE approach has a clearly defined tunable parameter—the cardinality of the exploration sequence—which can be adjusted according to the "hardness" (in terms of learning) of the reward distributions and observation models. It is thus more easily extendable to handle variations of MAB, including decentralized MAB with multiple players and incomplete reward observations under collisions, MAB with unknown Markov dynamics, and combinatorial MAB with depen-

dent arms that often arise in network optimization problems such as the shortest path, the minimum spanning tree, and the dominating set problems under unknown random weights.

Consider first a decentralized MAB problem in which multiple distributed players learn from their local observations and make decisions independently. While other players' observations and actions are unobservable, players' actions affect each other: conflicts occur when multiple players choose the same arm at the same time and conflicting players can only share the reward offered by the arm, not necessarily with conservation. Such an event is referred to as a collision and is unobservable to the players. In other words, a player does not know whether it is involved in a collision, or equivalently, whether the received reward reflects the true state of the arm. Collisions thus not only result in immediate reward loss, but also corrupt the observations that a player relies on for learning the arm rank. Such decentralized learning problems arise in communication networks where multiple distributed users share the access to a common set of channels, each with unknown communication quality. If multiple users access the same channel at the same time, no one transmits successfully or only one captures the channel through certain signaling schemes such as carrier sensing. Another application is multi-agent systems in which M agents search or collect targets in N locations. When multiple agents choose the same location, they share the reward in an unknown way that may depend on which player comes first or the number of colliding agents.

The deterministic separation of exploration and exploitation in DSEE, however, can ensure that collisions are contained within the exploitation sequence. Learning in the exploration sequence is thus carried out using only reliable ob-

servations. In particular, we show that under the DSEE approach, the system regret, defined as the total reward loss with respect to the ideal scenario of known reward models and centralized scheduling among players, grows at the same orders as the regret in the single-player MAB under the same conditions on the reward distributions. These results hinge on the extendability of DSEE to targeting at arms with arbitrary ranks (not necessarily the best arm) and the sufficiency in learning the arm rank solely through the observations from the exploration sequence.

3.1 The DSEE Policy

In this section, we present the DSEE approach and analyze its performance for both Sub-Gaussian and heavy-tailed reward distributions.

3.1.1 The General Structure

Time is divided into two interleaving sequences: an exploration sequence and an exploitation sequence. In the exploration sequence, the player plays all arms in a round-robin fashion. In the exploitation sequence, the player plays the arm with the largest sample mean (or a properly chosen mean estimator) calculated from past reward observations. It is also possible to use only the observations obtained in the exploration sequence in computing the sample mean. This leads to the same regret order with a significantly lower complexity since the sample mean of each arm only needs to be updated at the same sublinear rate as the exploration sequence. A detailed implementation of DSEE is given in Fig. 3.1.

The DSEE Approach

- Notations and Inputs: Let $\mathcal{A}(t)$ denote the set of time indices that belong to the exploration sequence up to (and including) time t. Let $|\mathcal{A}(t)|$ denote the cardinality of $\mathcal{A}(t)$. Let $\overline{\mu}_k(t)$ denote the sample mean of arm k computed from the reward observations at times in $\mathcal{A}(t-1)$. For two positive integers k and l, define $k \oslash l := ((k-1) \bmod l) + 1$, which is an integer taking values from $1, 2, \cdots, l$.
- At time t,
  1. if $t \in \mathcal{A}(t)$, play arm $k = |\mathcal{A}(t)| \oslash K$;
  2. if $t \notin \mathcal{A}(t)$, play arm $k^* = \arg\max\{\overline{\mu}_k(t),\ 1 \le k \le K\}$.

Figure 3.1: The DSEE approach for the classic MAB.

In DSEE, the tradeoff between exploration and exploitation is balanced by choosing the cardinality of the exploration sequence. To minimize the regret growth rate, the cardinality of the exploration sequence should be set to the minimum that ensures a reward loss in the exploitation sequence having an order no larger than the cardinality of the exploration sequence. The detailed regret analysis is given next. The analysis of DSEE is given under three general distribution model. Specifically, we show the following results for DSEE.

- Logarithmic regret under Sub-Gaussian distribution model.
- Sublinear regret under heavy-tailed distribution model with no prior knowledge.
- Logarithmic regret under heavy-tailed distribution model using truncated sample mean.

Before moving to the analysis of DSEE, we review the concentration results

on the sample mean of the distributions. Theses results are used in the analysis of the learning policies.

3.1.2 Preliminaries

Concentration inequalities are a key element in the analysis of the learning policies. In this section we briefly review two concentration results on Sub-Gaussian and heavy-tailed distributions. These results are used in the analysis throughout the dissertation.

Recall that a real-valued random variable X is called Sub-Gaussian if it satisfies the following [29],

$$\mathbb{E}[e^{u(X-\mathbb{E}[X])}] \leq e^{\zeta u^2/2} \tag{3.1}$$

for some constant $\zeta > 0$. When the distribution of the observations is Sub-Gaussian, Hoeffding-like concentration inequalities hold and it is easy to obtain the confidence intervals. Let X be a Sub-Gaussian random variable with mean μ. Let $\overline{X}(s)$ be the sample mean obtained from s i.i.d. observations of X. For any probability $p \in (0,1)$, [30]

$$\begin{aligned} &\mathbb{P}[\overline{X}(s) + \sqrt{\tfrac{2\zeta \log \frac{1}{p}}{s}} < \mu] \leq p, \\ &\mathbb{P}[\overline{X}(s) - \sqrt{\tfrac{2\zeta \log \frac{1}{p}}{s}} > \mu] \leq p. \end{aligned} \tag{3.2}$$

Equivalently,

$$\mathbb{P}[\overline{X}(s) - \mu < -\delta] \leq \exp(-as\delta^2),$$

$$\mathbb{P}[\overline{X}(s) - \mu > \delta] \leq \exp(-as\delta^2), \tag{3.3}$$

where $a = \frac{1}{2\xi}$ is a constant.

We also study the heavy-tailed distribution models. For heavy-tailed distributions, upper bounds on moment generating function (similar to 3.1) no longer exist. However, with the assumption of an upper bound on the moments of order $1 < b \leq 2$ we still can obtain similar confidence intervals. In particular, if for a random variable X

$$\mathbb{E}[X^b] \leq u, \tag{3.4}$$

for some $u > 0$, we can use the truncated sample mean defined as

$$\widehat{X}(s,p) = \frac{1}{s}\sum_{t=1}^{s} X(t)\mathbb{1}\left\{|X(t)| \leq (\frac{ut}{\log\frac{1}{p}})^{1/b}\right\} \tag{3.5}$$

to obtain confidence intervals on mean value of the observations. Particularly for any $p \in (0, \frac{1}{2}]$,

$$\begin{aligned}\Pr\left[\widehat{X}(s,p) - 4u^{1/b}(\tfrac{\log\frac{1}{p}}{s})^{\frac{b-1}{b}} > \mu\right] &\leq p \\ \Pr\left[\widehat{X}(s,p) + 4u^{1/b}(\tfrac{\log\frac{1}{p}}{s})^{\frac{b-1}{b}} < \mu\right] &\leq p.\end{aligned} \tag{3.6}$$

For a proof for 10.5, see Lemma 1 in [31].

3.1.3 Under Sub-Gaussian Reward Distributions

Under Sub-Gaussian reward distribution, the cardinality of the exploration sequence for DSEE policy is designed to be in $O(\log t)$ as specified in the following

theorem. We show in the following theorem that DSEE achieves the optimal logarithmic regret order for all Sub-Gaussian reward distributions.

Let σ be a permutation of $\{1, ..., K\}$ such that $\mu_{\sigma(1)} \geq \mu_{\sigma(2)} \geq \cdots \geq \mu_{\sigma(K)}$ and Δ_k^σ be the gap in the mean values of arm $\sigma(k)$ and arm $\sigma(1)$ (i.e. $*$). We might drop the superscript σ from Δ_k^σ when it is clear from the context.

Theorem 1 *Construct an exploration sequence as follows. Let a, ξ be the constants such that 3.3 holds. Define*[1] $\Delta_k \stackrel{\Delta}{=} \mu_{\sigma(1)} - \mu_{\sigma(k)}$ *for $k = 2, \ldots, K$. Choose a constant $c \in (0, \Delta_2)$, a constant $\delta = c/2$, and a constant $w > \frac{1}{a\delta^2}$. For each $t > 1$, if $|\mathcal{A}(t-1)| < K\lceil w \log t \rceil$, then include t in $\mathcal{A}(t)$. Under this exploration sequence, the resulting DSEE policy π^* has regret, $\forall T$,*

$$R_{\pi^*}(T; \mathcal{F}) \leq \sum_{k=2}^{K} \lceil w \log T \rceil \Delta_k + 2K\Delta_K (1 + \frac{1}{a\delta^2 w - 1}). \tag{3.7}$$

proof 1 *Let $R_{T,O}^{\pi^*}(\mathcal{F})$ and $R_{T,I}^{\pi^*}(\mathcal{F})$ denote, respectively, regret incurred during the exploration and the exploitation sequences. From the construction of the exploration sequence, it is easy to see that*

$$R_{\pi^*,O}(T; \mathcal{F}) \leq \sum_{k=2}^{K} \lceil w \log T \rceil \Delta_k. \tag{3.8}$$

During the exploitation sequence, a reward loss happens if the player incorrectly identifies the best arm. We thus have

$$\begin{aligned} R_{\pi^*,I}(T, \mathcal{F}) &\leq \mathbb{E}[\sum_{t \notin \mathcal{A}(T), t \leq T} \mathbb{I}(\pi^*(t) \neq \sigma(1))]\Delta_K \\ &= \sum_{t \notin \mathcal{A}(T), t \leq T} \Pr(\pi^*(t) \neq \sigma(1))\Delta_K. \end{aligned} \tag{3.9}$$

[1]Without loss of generality, we assume that $\{\mu_k\}_{k=1}^K$ are distinct.

For $t \notin \mathcal{A}(T)$, define the following event

$$\mathcal{E}(t) \triangleq \{|\overline{\mu}_k(t) - \mu_k| \leq \delta, \ \forall\ 1 \leq k \leq K\}. \tag{3.10}$$

From the choice of δ, it is easy to see that under $\mathcal{E}(t)$, the best arm is correctly identified. We thus have

$$\begin{aligned} R_{\pi^*,I}(T;\mathcal{F}) &\leq \sum_{t\notin\mathcal{A}(T),t\leq T} \Pr(\overline{\mathcal{E}(t)})\Delta_K \\ &= \sum_{t\notin\mathcal{A}(T),t\leq T} \Pr(\exists\ 1 \leq k \leq K\ s.t.\ |\overline{\mu}_k(t) - \mu_k| > \delta)\Delta_K \\ &\leq \sum_{t\notin\mathcal{A}(T),t\leq T} \sum_{k=1}^{K} \Pr(|\overline{\mu}_k(t) - \mu_k| > \delta)\Delta_K, \end{aligned} \tag{3.11}$$

where 3.11 results from the union bound. Recall that $\tau_k(t)$ denotes the number of times that arm k has been played during the exploration sequence up to time t. Applying 3.3 to 3.11, we have

$$\begin{aligned} R_{\pi^*,I}(T,\mathcal{F}) &\leq 2\Delta_K \sum_{t\notin\mathcal{A}(T),t\leq T} \Sigma_{k=1}^{K} \exp(-a\delta^2\tau_k(t)) \\ &\leq 2\Delta_K \sum_{t\notin\mathcal{A}(T),t\leq T} \Sigma_{k=1}^{K} \exp(-a\delta^2 w\log t) \qquad (3.12) \\ &= 2\Delta_K \sum_{t\notin\mathcal{A}(T),t\leq T} \sum_{k=1}^{K} t^{-a\delta^2 w} \\ &\leq 2K\Delta_K \Sigma_{t=1}^{\infty} t^{-a\delta^2 w} \\ &\leq 2K\Delta_K (1 + \frac{1}{a\delta^2 w - 1}), \qquad (3.13) \end{aligned}$$

where 3.12 comes from $\tau_k(t) \geq w\log t$ and 3.13 from $a\delta^2 w > 1$.

Combining 3.8 and 3.13, we arrive at the theorem.

The choice of the exploration sequence given in Theorem 1 is not unique. In particular, when the horizon length T is given, we can choose a single block of exploration followed by a single block of exploitation. In the case of infinite

horizon, we can follow the standard technique of partitioning the time horizon into epochs with geometrically growing lengths and applying the finite-T scheme to each epoch.

We point out that the logarithmic regret order requires certain knowledge about the differentiability of the best arm. Specifically, we need a lower bound (parameter c defined in Theorem 1) on the difference in the reward mean of the best and the second best arms. We also need to know the bounds on parameter ξ such that the Chernoff-Hoeffding bound 3.3 holds. These bounds are required in defining w that specifies the minimum leading constant of the logarithmic cardinality of the exploration sequence necessary for identifying the best arm. However, we show that when no knowledge on the reward models is available, we can increase the cardinality of the exploration sequence of π^* by an arbitrarily small amount to achieve a regret arbitrarily close to the logarithmic order.

Theorem 2 *Let $f(t)$ be any positive increasing sequence with $f(t) \to \infty$ as $t \to \infty$. Construct an exploration sequence as follows. For each $t > 1$, include t in $\mathcal{A}(t)$ if $|\mathcal{A}(t-1)| < K\lceil f(t)\log t\rceil$. The resulting DSEE policy π^* has regret*

$$R_{\pi^*}(T;\mathcal{F}) = O(f(T)\log T).$$

proof 2 *Recall constants a and δ defined in Theorem 1. Note that since $f(t) \to \infty$ as $t \to \infty$, there exists a t_0 such that for any $t > t_0$, $a\delta^2 f(t) \geq b$ for some $b > 1$. Similar to the proof of Theorem 1, we have, following 3.11,*

$$\begin{aligned} R_{\pi^*,I}(T;\mathcal{F}) &\leq 2K\Delta_K \sum_{t\notin\mathcal{A}(T),t\leq T} \exp(-a\delta^2 f(t)\log t) \\ &\leq 2K\Delta_K \sum_{t=1}^{t_0} \exp(-a\delta^2 f(t)\log t) + \sum_{t=t_0+1}^{\infty} t^{-b} \\ &\leq 2K\Delta_K(t_0 + \frac{1}{b-1}t_0^{1-b}). \end{aligned} \tag{3.14}$$

It is easy to see that

$$R_{\pi^*,O}(T;\mathcal{F}) \leq \sum_{n=2}^{N} \lceil f(T)\log T \rceil \Delta_n. \tag{3.15}$$

Combining 3.14 and 3.15, we have

$$R_{\pi^*}(T;\mathcal{F}) \leq \sum_{n=2}^{N} \lceil f(T)\log T \rceil \Delta_k \; + \; 2K\Delta_K(t_0 + \frac{1}{b-1}t_0^{1-b}). \tag{3.16}$$

From the proof of Theorem 2, we observe a tradeoff between the regret order and the finite-time performance. While one can arbitrarily approach the logarithmic regret order by reducing the diverging rate of $f(t)$, the price is a larger additive constant as shown in 3.16.

3.1.4 Sublinear Regret under Heavy-Tailed Distribution Model with Sublinear Complexity and No Prior Knowledge

For heavy-tailed reward distributions, the Chernoff-Hoeffding bound does not hold in general. A weaker bound on the deviation of the sample mean from the true mean is established in the lemma below.

Lemma 1 *Let $\{X(t)\}_{t=1}^{\infty}$ be i.i.d. random variables drawn from a distribution with finite pth moment ($p > 1$). Let $\overline{X}_t = \frac{1}{t}\Sigma_{s=1}^{t} X(s)$ and $\mu = \mathbb{E}[X(s)]$. We have, for all $\delta > 0$,*

$$\Pr(|\overline{X}_t - \theta| \geq \delta) \leq \begin{cases} (3\sqrt{2})^p p^{p/2} \frac{\mathbb{E}[|X(1)-\theta|^p]}{\delta^p} t^{1-p} & if\ p \leq 2 \\ (3\sqrt{2})^p p^{p/2} \frac{\mathbb{E}[|X(1)-\theta|^p]}{\delta^p} t^{-p/2} & if\ p > 2 \end{cases}$$

proof 3 *By Chebyshev's inequality we have,*

$$\Pr(|\overline{X}_t - \mu| > \delta) \quad \leq \quad \frac{\mathbb{E}[|\overline{X}_t - \mu|^p]}{\delta^p}$$

$$= \frac{\mathbb{E}[|\Sigma_{k=1}^{t}(X(k)-\theta)|^p]}{t^p\delta^p}$$
$$\leq B_p \frac{\mathbb{E}[(\Sigma_{k=1}^{t}(X(k)-\theta)^2)^{p/2}]}{t^p\delta^p}, \quad (3.17)$$

where 3.17 holds by the Marcinkiewicz-Zygmund inequality for some B_p depending only on p. The best constant in the Marcinkiewicz-Zygmund inequality was shown in [32] to be $B_p \leq (3\sqrt{2})^p p^{p/2}$.

Next, we prove Lemma 1 by considering the two cases of p.

- *$p \leq 2$: Considering the inequality $(\Sigma_{s=1}^{t} a_s)^\alpha \leq \Sigma_{s=1}^{t} a_s^\alpha$ for $a_s \geq 0$ and $\alpha \leq 1$ (which can be easily shown using induction), we have, from 3.17,*

$$\Pr(|\overline{X}_t - \mu| \geq \delta) \leq B_p \frac{\mathbb{E}[\Sigma_{s=1}^{t}|X(s)-\mu|^p]}{t^p\delta^p}$$
$$= B_p \frac{\mathbb{E}[|X(1)-\mu|^p]}{\delta^p} t^{1-p}. \quad (3.18)$$

- *$p > 2$: Using Jensen's inequality, we have, from 3.17,*

$$\Pr(|\overline{X}_t - \theta| \geq \delta) \leq B_p \frac{\mathbb{E}[t^{p/2-1}\Sigma_{s=1}^{t}|X(s)-\mu|^p]}{t^p\delta^p}$$
$$= B_p \frac{\mathbb{E}[|X(1)-\mu|^p]}{\delta^p} t^{-p/2}. \quad (3.19)$$

Based on Lemma 1, we have the following results on the regret performance of DSEE under heavy-tailed reward distributions.

Theorem 3 *Assume that the reward distributions have finite pth order moment ($p > 1$). Construct an exploration sequence as follows. Choose a constant $v > 0$. For each $t > 1$, include t in $\mathcal{A}(t)$ if $|\mathcal{A}(t-1)| < vt^{1/p}$ for $1 < p \leq 2$ or $|\mathcal{A}(t-1)| < vt^{\frac{1}{1+p/2}}$ for $p > 2$. Under this exploration sequence, the resulting DSEE policy π^p has regret*

$$R_{\pi^p}(T;\mathcal{F}) = \begin{cases} O(T^{1/p}) & if \ \ 1 < p \leq 2 \\ O(T^{\frac{1}{1+p/2}}) & if \ \ p > 2 \end{cases} \quad (3.20)$$

An upper bound on the regret for each T is given in (3.21) in the proof.

proof 4 *We prove the theorem for the case of $p > 2$, the other case can be shown similarly. Following a similar line of arguments as in the proof of Theorem 1, we can show, by applying Lemma 1 to 3.11,*

$$\begin{aligned} R_{\pi^p,I}(T;\mathcal{F}) &\leq \Delta_K B_p \frac{\mathbb{E}[|X(1)-\mu|^p]}{\delta^p} v^{-p/2} \sum_{t=1}^{T} t^{\frac{-p/2}{1+p/2}} \\ &\leq \Delta_K B_p \frac{\mathbb{E}[|X(1)-\mu|^p]}{\delta^p} v^{-p/2} [(1+p/2)(T^{\frac{1}{1+p/2}}-1)+1] \end{aligned}$$

Considering the cardinality of the exploration sequence, we have, $\forall T$,

$$R_{\pi^p}(T;\mathcal{F}) \leq \begin{cases} \Delta_N B_p \frac{\mathbb{E}[|X(1)-\theta|^p]}{(\Delta_2/2)^p} v^{-p/2} [p(T^{\frac{1}{p}}-1)+1] & \\ +\Delta_N \lceil vT^{\frac{1}{p}} \rceil & if \;\; p \leq 2 \\ \Delta_N B_p \frac{\mathbb{E}[|X(1)-\theta|^p]}{(\Delta_2/2)^p} v^{-p/2} [(1+p/2)(T^{\frac{1}{1+p/2}}-1)+1] & \\ +\Delta_N \lceil vT^{\frac{1}{1+p/2}} \rceil & if \;\; p > 2 \end{cases} \tag{3.21}$$

The regret order given in Theorem 3 is thus readily seen.

3.1.5 Logarithmic Regret under Heavy-Tailed Distribution Model using Truncated Sample Mean

Inspired by Bubeck, Cesa-Bianchi, and Lugosi's work [31], we show that using the truncated sample mean, DSEE can offer logarithmic regret order for heavy-tailed reward distributions with a carefully chosen cardinality of the exploration

sequence. Similar to the UCB variation developed in [31], this logarithmic regret order is achieved at the price of prior information on the reward distributions and higher computational and memory requirement. The computational and memory requirement, however, is significantly lower than that of the UCB variation in [31], since the DSEE approach only needs to store samples from and compute the truncated sample mean at the exploration times with $O(\log T)$ order rather than each time instant.

Theorem 4 *Assume that the reward of each arm satisfies* $\mathbb{E}[|X_k(1)|^p] \leq u$ *for some constants* $u > 0$ *and* $p \in (1, 2]$. *Let* $a = 4^{\frac{p}{1-p}} u^{\frac{1}{1-p}}$. *Define* $\Delta_k \stackrel{\Delta}{=} \mu_{\sigma(1)} - \mu_{\sigma(k)}$ *for* $k = 2, \ldots, K$. *Construct an exploration sequence as follows. Choose a constant* $\delta \in (0, \Delta_2/2)$ *and a constant* $w > \frac{1}{a\delta^{p/(p-1)}}$. *For each* $t > 1$, *if* $|\mathcal{A}(t-1)| < K\lceil w \log t \rceil$, *then include* t *in* $\mathcal{A}(t)$. *At an exploitation time* t, *play the arm with the largest truncated sample mean given by*

$$\widehat{\mu}_k(\tau_k(t), \epsilon_k(t)) = \frac{1}{\tau_k(t)} \sum_{s=1}^{\tau_k(t)} X_{k,s} \mathbb{1}\{|X_k(s)| \leq (\frac{us}{\log(\epsilon_k(t)^{-1})})^{1/p}\},$$

where $X_k(s)$ *denotes the* s*th observation of arm* k *during the exploration sequence,* $\tau_k(t)$ *the total number of such observations, and* $\epsilon_k(t)$ *in the truncation for each arm at each time* t *is given by*

$$\epsilon_k(t) = \exp(-a\delta^{\frac{p}{p-1}} \tau_k(t)). \tag{3.22}$$

The resulting DSEE policy π^* *has regret*

$$R_{\pi^*}(T; \mathcal{F}) \leq \sum_{k=2}^{K} \lceil w \log T \rceil \Delta_k + 2K\Delta_K (1 + \frac{1}{a\delta^{p/(p-1)} w - 1}). \tag{3.23}$$

proof 5 *Following the same line of arguments as in the proof of Theorem 1, we have, following 3.11*

$$R_{\pi^*, I}(T; \mathcal{F}) \leq \sum_{t \notin \mathcal{A}(T), t \leq T} \sum_{k=1}^{K} \Pr(|\widehat{\mu}_k(\tau_k(t), \epsilon_k(t)) - \mu_k| > \delta) \Delta_K. \tag{3.24}$$

Based on 10.5, we have,

$$\Pr(|\hat{\mu}(\tau_k(t), \epsilon_k(t)) - \mu_k| > \delta) \leq 2\exp(-a\delta^{\frac{p}{p-1}}\tau_k(t)). \tag{3.25}$$

Substituting the above equation into 3.24, we have

$$\begin{aligned} R_{\pi^*,I}(T;\mathcal{F}) &\leq 2\Delta_K \sum_{t\notin\mathcal{A}(T),t\leq T} \Sigma_{k=1}^{K} \exp(-a\delta^{\frac{p}{p-1}}\tau_k(t)) \\ &\leq 2\Delta_K \sum_{t\notin A(T),t\leq T} \Sigma_{k=1}^{K} \exp(-a\delta^{\frac{p}{p-1}} w\log t) \\ &= 2\Delta_K \sum_{t\notin A(T),t\leq T} \sum_{k=1}^{K} t^{-a\delta^{p/(p-1)}w} \\ &\leq 2K\Delta_K \sum_{t=1}^{\infty} t^{-a\delta^{p/(p-1)}w} \\ &\leq 2K\Delta_K(1+\frac{1}{a\delta^{p/(p-1)}w-1}) \end{aligned} \tag{3.26}$$

We then arrive at the theorem, considering $R_{T,O}^{\pi^*}(\mathcal{F}) \leq \Sigma_{k=2}^{K}\lceil w\log T\rceil\Delta_k$.

We point out that to achieve the logarithmic regret order under heavy-tailed distributions, an upper bound on $\mathbb{E}[|X_n(1)|^p]$ for a certain p needs to be known. The range constraint of $p \in (1,2]$ in Theorem 4 can be easily addressed: if we know $\mathbb{E}[|X_n(1)|^p] \leq u$ for a certain $p > 2$, then $\mathbb{E}[|X|^2] \leq u+1$. Similar to Theorem 2, we can show that when no knowledge on the reward models is available, we can increase the cardinality of the exploration sequence by an arbitrarily small amount (any diverging sequence $f(t)$) to achieve a regret arbitrarily close to the logarithmic order. One necessary change to the policy is that the constant δ in Theorem 4 used in 3.22 for calculating the truncated sample mean should be replaced by $f(t)^\gamma$ for some $\gamma \in (\frac{1-p}{p}, 0)$.

3.2 Extendibility to Variations of MAB

In this section, we extend the DSEE approach to several MAB variations including MAB with various objectives, decentralized MAB with multiple players and incomplete reward observations under collisions. Extensions to restless MAB with unknown dynamics and combinatorial MAB with dependent arms can be found in [33], [34].

3.2.1 MAB under Various Objectives

Consider a generalized MAB problem in which the desired arm is the mth best arm for an arbitrary $1 \leq m \leq K$. Such objectives may arise when there are multiple players (see the next subsection) or other constraints/costs in arm selection. The classic policies in [1, 2, 3] cannot be directly extended to handle this new objective. For example, for the UCB policy proposed by Auer *et al.* in [3], simply choosing the arm with the mth largest index cannot guarantee an optimal solution. This can be seen from the index form given in 2.5: when the index of the desired arm is too large to be selected, its index tends to become even larger due to the second term of the index. The rectification proposed in [35] is to combine the upper confidence bound with a symmetric lower confidence bound. Specifically, the arm selection is completed in two steps at each time: the upper confidence bound is first used to filter out arms with a lower rank, the lower confidence bound is then used to filter out arms with a higher rank. It was shown in [35] that under the extended UCB, the expected time that the player does not play the targeted arm has a logarithmic order.

The DSEE approach, however, can be directly extended to handle this general objective. Under DSEE, all arms, regardless of their ranks, are sufficiently explored by carefully choosing the cardinality of the exploration sequence. As a consequence, this general objective can be achieved by simply choosing the arm with the mth largest sample mean in the exploitation sequence. Specifically, assume that a cost $C_j > 0$ ($j \neq m, 1 \leq j \leq K$) is incurred when the player plays the jth best arm. Define the regret $R_T^\pi(\mathcal{F}; m)$ as the expected total costs over time T under policy π.

Theorem 5 *By choosing the parameter c in Theorem 1 to satisfy $0 < c < \min\{\Delta_m - \Delta_{m-1}, \Delta_{m+1} - \Delta_m\}$ or a parameter δ in theorem 3 and 4 to satisfy $0 < \delta < \frac{1}{2}\min\{\Delta_m - \Delta_{m-1}, \Delta_{m+1} - \Delta_m\}$ and letting the player select the arm with the m-th largest sample mean (or truncated sample mean in case of 4) in the exploitation sequence, Theorems 1-4 hold for $R_T^\pi(\mathcal{F}; m)$.*

proof 6 *The proof is similar to those of previous theorems. The key observation is that after playing all arms sufficient times during the exploration sequence, the probability that the sample mean of each arm deviates from its true mean by an amount larger than the non-overlapping neighbor is small enough to ensure a properly bounded regret incurred in the exploitation sequence.*

We now consider an alternative scenario that the player targets at a set of best arms, say the M best arms. We assume that a cost is incurred whenever the player plays an arm not in the set. Similarly, we define the regret $R_T^\pi(\mathcal{F}; M)$ as the expected total costs over time T under policy π.

Theorem 6 *By choosing the parameter c in Theorem 1 to satisfy* $0 < c < \Delta_{M+1} - \Delta_M$ *or a parameter* δ *in theorem 3 and 4 to satisfy* $0 < \delta < \frac{1}{2}(\Delta_{M+1} - \Delta_M)$ *and letting the player select one of the M arms with the largest sample means (or truncated sample mean in case of 4) in the exploitation sequence, Theorem 1-4 hold for* $R_T^{\pi}(\mathcal{F}; M)$.

proof 7 *The proof is similar to those of previous theorems. Compared to Theorem 5, the condition on c for applying Theorem 1 is more relaxed: we only need to know a lower bound on the mean difference between the M-th best and the* $(M+1)$*-th best arms. This is due to the fact that we only need to distinguish the M best arms from others instead of specifying their rank.*

By selecting arms with different ranks of the sample mean in the exploitation sequence, it is not difficult to see that Theorem 5 and Theorem 6 can be applied to cases with time-varying objectives.

In the next subsection, we use these extensions of DSEE to solve a class of decentralized MAB with incomplete reward observations.

3.2.2 Decentralized MAB with Incomplete Reward Observations

Distributed Learning under Incomplete Observations

Consider M distributed players. At each time t, each player chooses one arm to play. When multiple players choose the same arm (say, arm K) to play at time t, a player (say, player m) involved in this collision obtains a potentially

reduced reward $Y_{k,m}(t)$ with $\sum_{m=1}^{M} Y_{k,m}(t) \leq X_k(t)$. We focus on the case where the M best arms have positive reward mean and collisions cause reward loss. The distribution of the partial reward $Y_{k,m}(t)$ under collisions can take any unknown form and has any dependency on k, m and t. Players make decisions solely based on their partial reward observations $Y_{k,m}(t)$ without information exchange. Consequently, a player does not know whether it is involved in a collision, or equivalently, whether the received reward reflects the true state $X_k(t)$ of the arm.

A local arm selection policy π_m of player m is a function that maps from the player's observation and decision history to the arm to play. A decentralized arm selection policy π is thus given by the concatenation of the local polices of all players:

$$\pi_d \stackrel{\Delta}{=} [\pi_1, \cdots, \pi_M].$$

The system performance under policy π_d is measured by the system regret $R_T^{\pi_d}(\mathcal{F})$ defined as the expected total reward loss up to time T under policy π_d compared to the ideal scenario that players are centralized and $\mathcal{F}$ is known to all players (thus the M best arms with highest means are played at each time). We have

$$R_T^{\pi_d}(\mathcal{F}) \stackrel{\Delta}{=} T \sum_{m=1}^{M} \theta_{\sigma(m)} - \mathbb{E}[\Sigma_{t=1}^{T} Y_{\pi_d}(t)],$$

where $Y_{\pi_d}(t)$ is the total random reward obtained at time t under decentralized policy π_d. Similar to the single-player case, any policy with a sublinear order of regret would achieve the maximum average reward given by the sum of the M highest reward means.

Decentralized Policies under DSEE

In order to minimize the system regret, it is crucial that each player extracts reliable information for learning the arm rank. This requires that each player obtains and recognizes sufficient observations that were received without collisions. As shown in Sec. 3.1, efficient learning can be achieved in DSEE by solely utilizing the observations from the deterministic exploration sequence. Based on this property, a decentralized arm selection policy can be constructed as follows. In the exploration sequence, players play all arms in a round-robin fashion with different offsets which can be predetermined based on, for example, the players' IDs, to eliminate collisions. In the exploitation sequence, each player plays the M arms with the largest sample mean calculated using only observations from the exploration sequence under either a prioritized or a fair sharing scheme. While collisions still occur in the exploitation sequences due to the difference in the estimated arm rank across players caused by the randomness of the sample means, their effect on the total reward can be limited through a carefully designed cardinality of the exploration sequence. Note that under a prioritized scheme, each player needs to learn the specific rank of one or multiple of the M best arms and Theorem 5 can be applied. While under a fair sharing scheme, a player only needs to learn the set of the M best arms (as addressed in Theorem 6) and use the common arm index for fair sharing. An example based on a round-robin fair sharing scheme is illustrated in Fig. 3.2. We point out that under a fair sharing scheme, each player achieves the same average reward at the same rate.

Theorem 7 *Under a decentralized policy based on DSEE, Theorem 1-4 hold for* $R_T^{\pi_d}(\mathcal{F})$.

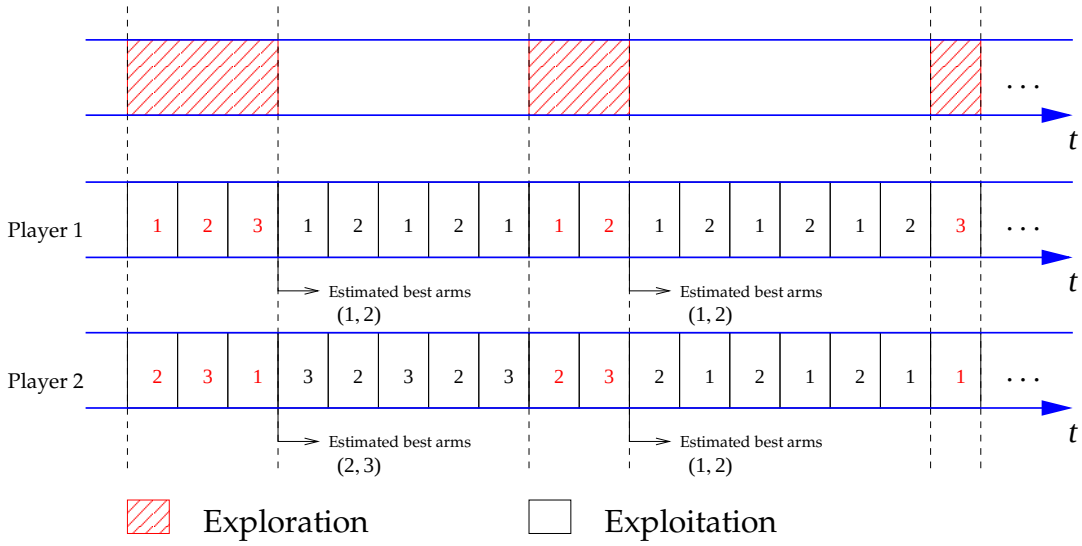

Figure 3.2: An example of decentralized policies based on DSEE ($M = 2$, $K = 3$, the index of the selected arm at each time is given).

proof 8 *The regret in the decentralized policy is completely determined by the learning efficiency of the M best arms at each player. The key is to notice that during the exploitation sequence, collisions can only happen if at least one player incorrectly identifies the M best arms. As a consequence, to analyze the regret in the exploitation sequence, we only need to consider such events. The proof is thus similar to those of previous theorems.*

3.3 Conclusion

The DSEE approach addresses the fundamental tradeoff between exploration and exploitation in MAB by separating, in time, the two often conflicting objectives. It has a clearly defined tunable parameter—the cardinality of the exploration sequence—which can be adjusted to handle any reward distributions and the lack of any prior knowledge on the reward models. Furthermore, the

deterministic separation of exploration from exploitation allows easy extensions to variations of MAB, including decentralized MAB with multiple players and incomplete reward observations under collisions, MAB with unknown Markov dynamics, and combinatorial MAB with dependent arms that often arise in network optimization problems such as the shortest path, the minimum spanning tree, and the dominating set problems under unknown random weights.

In algorithm design, there is often a tension between performance and generality. The generality of the DSEE approach comes at a price of finite-time performance. Even though DSEE offers the optimal regret order for any distribution, simulations show that the leading constant in the regret offered by DSEE is often inferior to that of classic policies proposed in [1, 2, 3] that target at specific types of distributions.

CHAPTER 4

TIME-VARYING STOCHASTIC MULTI-ARMED BANDIT PROBLEMS

In this chapter, we address time variation in the reward processes in a MAB problem. We first consider the case where there is no restriction on how often the reward distribution of each arm can change. We adopt the performance measure of weak regret first introduced in [12] for non-stochastic MAB problems. Weak regret is defined as the total expected reward loss over a time horizon of length T when compared to the optimal single-arm policy with full knowledge of the reward models. In other words, the performance of the player is measured against a genie who knows, non-causally, the entire reward distribution sequence of every arm but is restricted to play a fixed arm. Note that with arbitrary variations of the reward model, the performance measure of strict regret that allows the genie to arbitrarily switch among arms becomes meaningless: any arm selection policy would have linear regret order since past observations bear no information for current and future rewards due to the arbitrary variations in the reward models. In other words, it is impossible to approach the average performance of an omniscient genie bounded by no constraints. Weak regret, however, leads to a meaningful performance measure and admits tractable solutions. In this case, what the player is trying to learn is which arm has the largest cumulative expected reward rather than trying to catch the largest expected reward at each time instant. Intuitively, the former is possible as past reward observations become increasingly more informative for learning the largest cumulative reward as time goes.

By constructing a specific worst-case scenario where the mean values of Bernoulli distributed arms are approaching each other over time, we establish

that the weak regret of the stochastic MAB with arbitrary reward model variations is lower bounded by $\Theta(\sqrt{T})$. We show that this lower bound is achieved by an online learning algorithm, thus demonstrating the tightness of the lower bound and the order optimality of the algorithm.

We then consider a time variation model where the number of changes experienced by the reward distributions is constrained by a constant that is a function of the horizon length T. Referred to as the piece-wise stationary model, this time variation model allows reward distributions to change at arbitrary time instants, but the total number of change points is no more than $H(T) - 1$. In other words, the reward distribution sequence consists of up to $H(T)$ stationary segments. Under this time variation model, we show that an $O(\sqrt{TH(T)\log T})$ strict regret (where the genie has no constraint on arm selection) is achievable, which is sublinear in T provided that $H(T)$ is of a lower order than $\frac{T}{\log T}$. This model can also be interpreted as employing a more general definition of weak regret under the arbitrary variation model where the performance of a learning policy is measured against a genie who is allowed to switch arms no more than $H(T) - 1$ times over the time horizon of length$T$.

The time-varying stochastic MAB problem can also be seen as a variation of the classic non-stochastic MAB. In [12], Auer *et al.* considered the non-stochastic MAB problem where the rewards of each arm are given by a deterministic sequence assigned by an adversary. They introduced the concept of weak regret and established a $\Theta(\sqrt{T})$ lower bound on the *finite-time* weak regret where the online learning algorithms are measured by a worst-case scenario chosen by an adversary with the knowledge of the horizon length T. The time-varying stochastic MAB is similar to the non-stochastic problem except an arbitrary time-

varying distribution model is chosen by adversary instead of a deterministic reward sequence.

We point out that the $\Theta(\log T)$ lower bound on the model-specific regret of the classic stochastic MAB developed by Lai and Robbins in [1] is within an asymptotic regime where the worst-case is independent of the time horizon length T. While, $\Theta(\sqrt{T})$ lower bound on the model-independent regret for the classic MAB is obtained by a worst case reward distribution model that depends on the time horizon T. Our result on the lower bound of the weak regret for the time-varying stochastic MAB problem is without knowledge of T. The significance of this result is better understood when compared to the classic stochastic MAB problem that admits two drastically different regret orders ($\Theta(\log T)$ and $\Theta(\sqrt{T})$).

There are several related results on time-varying stochastic MAB problems. A piece-wise stationary model was studied in [36] under the assumption that at each time, the rewards of a particular number of arms that are not played are known (the so-called side observations). It was shown in [36] that under specific assumptions on the difference between the mean-value of the best and the second best arms at each time instance and a minimum amount of changes in the mean value of at least one arm at each change point, logarithmic order regret is achievable. In [37], Garivier and Moulines studied the performance of two variations of the UCB algorithms under the same arbitrary time variation model considered in this work. They considered the regret definition where the genie is allowed to switch arms no more than $H(T) - 1$ times and showed that the two variations of the UCB algorithms (discounted UCB and sliding-window UCB) achieve $O(\sqrt{TH(T)}\log T)$ and $O(\sqrt{TH(T)\log T})$ regret performance, respective-

ly. They also established a $\Theta(\sqrt{T})$ lower bounded on regret where the genie is allowed to switch arms a constant number of times over the time horizon T. This lower bound, however, does not apply to the weak regret considered in this work where the genie is not allowed to switch arms. In [38], a specific time variation was considered in which the distribution sequences of the arms in a two-armed bandit approach each other over time. Within this set-up, a sufficient condition and a necessary condition were established to characterize when learning and tracking the best arm is possible.

4.1 Problem Formulation and Preliminaries

The rewards of each arm $i = 1, ..., K$ is characterized by a sequence of distributions $\mathcal{F}_i^{(T)} = \{f_{1,1}, f_{1,2}, ..., f_{1,T}\}$ chosen by an adversary. Define $\mathcal{F}^{(T)} = \{\mathcal{F}_1^{(T)}, \mathcal{F}_2^{(T)}, ..., \mathcal{F}_K^{(T)}\}$. We assume that all the distributions have bounded support. Without loss of generality, we assume that the bounded support is the unit interval $[0, 1]$. Let $X_i(t)$ denote the reward of arm i at time t, which is a random variable distritbuted according to $f_{i,t}$. Let $\mu_i(t) = \mathbb{E}[X_i(t)]$.

The player chooses one arm to play at each time t according to an arm selection policy π. The performance of a policy π is measured by regret, defined as

$$R_\pi(T) = \mathbb{E}[\sum_{t=1}^{T} X_{g(t)}(t) - \sum_{t=1}^{T} X_{\pi(t)}(t)], \tag{4.1}$$

where $g(t)$ is the optimal arm selected by a genie who knows the entire sequence of the reward distributions of each arm. Under the performance measure of weak regret, the genie is only allowed to play a fixed arm over the entire time

horizon. The weak regret can thus be written as

$$R_\pi^w(T) = \max_i \sum_{t=1}^{T} \mu_i(t) - \mathbb{E}[\sum_{t=1}^{T} X_{\pi(t)}(t)]. \tag{4.2}$$

4.2 Weak Regret

The Exp3 Algorithm [12]

- Parameter: $\gamma \in (0, 1]$
- Initialization: $w_i(1) = 1$ for $i = 1, ..., K$
- At time t,
  1. Set $p_i(t) = (1 - \gamma)\frac{w_i(t)}{\sum_{j=1}^{K} w_j(t)} + \frac{\gamma}{K}$
  2. Draw an arm $\pi(t)$ randomly with probabilities $p_1(t), ..., p_K(t)$
  3. For arm $\pi(t)$ set

$$w_{\pi(t)}(t + 1) = w_{\pi(t)}(t) \exp(\frac{\gamma X_{\pi(t)}(t)}{p_{\pi(t)} K})$$

  4. For other arms $(\{1, ..., K\} - \{\pi(t)\})$ set

$$w_j(t + 1) = w_j(t)$$

Figure 4.1: The EXP3 Algorithm

In this section, we show that for any policy π designed by the player, there exists a sequence of reward distributions such that the weak regret (defined in 4.2) grows at least as fast as $\Theta(\sqrt{T})$.

Theorem 8 *Consider a time-varying stochastic MAB problem with K ($K \geq 2$) arms. The weak regret of any policy π satisfies, for some constant $c > 0$ and $T_0 \in \mathbb{N}$,*

$$R_\pi^w(T) \geq c\sqrt{KT}, \tag{4.3}$$

for all $T > T_0$.

proof 9 *To prove this theorem, we consider a K-armed bandit with Bernoulli distributed rewards assigned to the arms. The reward distribution of one arm is Bernoulli with mean $\frac{1}{2} + \epsilon_t$ at time t. The reward distributions of all other arms are Bernoulli with mean $\frac{1}{2}$ for all t. The sequence ϵ_t diminishes to zero as t approaches to infinity. The diminishing rate of ϵ_t is carefully designed in order to obtain a tight lower bound on the weak regret. The detailed proof is given in Appendix A.*

In the proof of Theorem 8, the designed worst-case sequence of distributions is independent of the time horizon length T. Whether this lower bound gives the exact order of regret is answered by Theorem 23 below which provides an upper bound with the same order as the lower bound. Specifically, we show that the EXP3 algorithm introduced in [12] for the non-stochastic MAB directly applies to the time-varying stochastic MAB problem and offers an $O(\sqrt{T})$ regret order. The algorithm is described in Fig. 4.1.

Theorem 9 *For the time-varying stochastic MAB problem, with choice of parameter $\lambda = \sqrt{\frac{K \log K}{(e-1)T}}$, the EXP3 algorithm archives the weak regret performance*

$$R^w_{EXP3}(T) = O(\sqrt{KT \ln K}) \tag{4.4}$$

proof 10 *See Appendix A for the proof.*

We have assumed the time horizon T is known in advance to tune the parameter γ. This assumption is not essential. For a time-varying stochastic MAB problem with unknown time horizon we can partition the time horizon into epoches. Let $r = 0, 1, 2, ...$ denote the index of epoches. The length of epoch r is

set to $l_r = 2^r$. Restart the EXP3 algorithm at the beginning of each epoch with parameter $\gamma = \sqrt{\frac{K \log K}{(e-1)l_r}}$. A similar argument as in [12] shows that the resulting

The Exp3.S Algorithm [12]

- Parameters: $\gamma \in (0, 1]$ and $\alpha > 0$
- Initialization: $w_i(1) = 1$ for $i = 1, ..., K$
- At time t,
  1. Set $p_i(t) = (1 - \gamma)\frac{w_i(t)}{\sum_{j=1}^{K} w_j(t)} + \frac{\gamma}{K}$
  2. Draw an arm $\pi(t)$ randomly with probabilities $p_1(t), ..., p_K(t)$
  3. For arm $\pi(t)$ set

$$\begin{aligned} w_{\pi(t)}(t+1) &= w_{\pi(t)}(t) \exp\left(\frac{\gamma X_{\pi(t)}(t)}{p_{\pi(t)} K}\right) \\ &+ \frac{e\alpha}{K} \sum_{j=1}^{K} w_i(t) \end{aligned}$$

  4. For other arms $(\{1, ..., K\} - \{\pi(t)\})$ set

$$w_j(t+1) = w_j(t) + \frac{e\alpha}{K} \sum_{j=1}^{K} w_j(t)$$

Figure 4.2: The EXP3.S Algorithm

algorithm offers the same regret order.

4.3 Piece-wise Stationary Time Variation Model

In the arbitrary time variation model, since the past observations bear no information for current and future rewards, learning under a strict regret definition is infeasible. In this section, we consider a time variation model where the num-

ber of changes experienced by the reward distributions over a time horizon of length T is upper bounded by $H(T) - 1$. In other words, the reward distribution sequence consists of up to $H(T)$ stationary segments. This time variation model is thus referred to as the piece-wise stationary model. Under this model, we consider the strict regret where the genie has no constraint on arm selection.

$$R_\pi(T) = \sum_{t=1}^{T} \max_i \mu_i(t) - \mathbb{E}[\sum_{t=1}^{T} X_{\pi(t)}(t)]. \tag{4.5}$$

Note that this model can also be interpreted as employing a more general definition of weak regret under the arbitrary variation model where the performance of a learning policy is measured against a genie who is allowed to switch arms no more than $H(T) - 1$ times over the time horizon of length T.

The EXP3.S algorithm proposed in [12] for the non-stochastic MAB (see Fig. 4.2) applies to the stochastic MAB problem under the piece-wise stationary model and achieves $O(\sqrt{TH(T)\log T})$ regret as given in the theorem below.

Theorem 10 *In the piece-wise stationary time-varying stochastic MAB , the EXP3.S algorithm achives*

$$R_{EXP3.S}^{H(T)} = O(\sqrt{KTH(T)\log(KT)}). \tag{4.6}$$

proof 11 *See Appendix A for the proof.*

The length of time horizon T and and the value of $H(T)$ are used to determine the value of parameters α and γ in EXP3.S algorithm. If the time horizon is not known in advance, we can use the following epoch structure to obtain the same regret order. Partition the time horizon into epoches denoted by $r = 0, 1, 2, ...$ with length $l_r = 2^r$. Restart the EXP3.S algorithm at the beginning of each epoch

with parameters $\alpha = \frac{1}{l_r}$ and $\gamma = \sqrt{\frac{K(H(L_r)\log(Kl_r)+e)}{(e-1)l_r}}$. We show in the theorem below that this approach results in the same regret order for concave $H(T)$. The same result holds when only a concave order $\tilde{H}(T) = \Theta(H(T))$ of $H(T)$ (rather than $H(T)$) is known to the player. In this case, the parameter γ for each epoch can be chosen using $\tilde{H}(L_r)$ rather than $H(L_r)$.

Theorem 11 *The EXP3.S algorithm with the epoch structure described above achieves the following regret performance*

$$R_{EXP3.S}(T) = O(\sqrt{KTH(T)\log(KT)}). \tag{4.7}$$

proof 12 *See Appendix A for the proof.*

4.4 Conclusion

In this chapter we studied time-varying stochastic MAB problem under two models. In the first model the unknown reward distribution of each arm can change arbitrarily. In this arbitrary time variation model we obtained an $O(\sqrt{T})$ lower bound on weak regret. Moreover we showed that the EXP3 algorithm applies and obtains the optimal order weak regret. In the second model, which is a piece-wise stationary reward model, we showed that the EXP3.S algorithm applies and offers $O(\sqrt{TH(T)\log T})$ strict regret.

CHAPTER 5

RISK-AVERSE MULTI-ARMED BANDIT UNDER MEAN-VARIANCE MEASURE

The classic MAB formulation targets at maximizing the *expected* return of an online learning policy. In many applications, especially in economics and finance, a player may be more interested in reducing the uncertainty (i.e., risk) in the outcome, rather than achieving the highest ensemble average. The focus of this chapter is to develop results on risk-averse MAB, parallel to those on the classic risk-neutral MAB problems under the measure of mean-variance.

The notions of risk and uncertainty have been widely studied, especially in economics and mathematical finance. A commonly adopted risk measure is *mean-variance* [40]. Introduced by Markowits in 1952, mean-variance is particularly favored for portfolio selection in finance [41]. Specifically, the mean-variance $\xi(X)$ of a random variable X is given by

$$\xi(X) = \sigma^2(X) - \rho\,\mu(X), \tag{5.1}$$

where $\sigma^2(X)$ and $\mu(X)$ are, respectively, the variance and the mean of X, the coefficient $\rho > 0$ is the risk tolerance factor that balances the two objectives of high return and low risk. The definition of mean-variance can be interpreted as the Lagrangian relaxation of the constrained optimization problem of minimizing the risk (measured by the variance) for a given expected return or maximizing the expected return for a given level of risk.

In [42], a risk-averse MAB formulation based on the metric of mean-variance of observations was studied. Specifically, let $\pi(t)$ ($t = 1, 2, \ldots, T$), denote the arm played by a policy π and $X_{\pi(t)}(t)$ the observed reward at time t. The cumulative

mean-variance of the observed reward process is given by[1]

$$\xi_\pi(T) = \mathbb{E}\Big[\sum_{t=1}^{T} [(X_{\pi(t)}(t) - \frac{1}{T} \sum_{t=1}^{T} X_{\pi(t)}(t))^2 - \rho X_{\pi(t)}(t)] \Big], \tag{5.2}$$

where the first term inside the expectation corresponds to the cumulative empirical variance and the second term the cumulative empirical mean. The objective is a learning policy that minimizes $\xi_\pi(T)$. In this risk-averse model, the time variations in the observed reward process are considered as risk (we will discuss the motivating applications for this metric later). Similar to risk-neutral MAB, regret is defined as the performance loss with respect to the optimal policy under a known model.

While conceptually similar, regret in terms of mean-variance of observations differs from that in total expected reward in several major aspects that complicate the analysis of the lower bounds and algorithm performance. First, under the measure of expected reward, the optimal policy under a known model is to play the arm with the highest mean value over the entire horizon. Under the measure of mean-variance, however, the optimal policy in the known model case is not necessarily a single-arm policy (as shown in Sec. 5.4) and is in general intractable. Second, under the measure of mean-variance, regret can no longer be written as the sum of certain properly defined immediate performance loss at each time instant. More specifically, under the measure of mean-variance of observations, the contribution from playing a suboptimal arm at a given time t to the overall regret cannot be determined without knowing the entire sequence of decisions and observations. Third, regret in mean-variance involves higher order statistics of the random time spent on each arm. These fundamental differences in the behavior of regret are what render the problem difficult and call

[1]Notice that the cumulative mean-variance of the observed reward process is considered in contrast to a normalized version divided by T as considered in [42]. This definition facilitates the comparison with the risk-neutral MAB results given in Table 5.1.

for different techniques from that used in risk-neutral MAB problems.

The focus of [42] was on developing learning policies. Specifically, two learning policies were developed and analyzed. The first one is a variation of the UCB policy (referred to as MV-UCB), and the second a variation of the DSEE policy (referred to as MV-DSEE) originally developed in [3] and [4] for risk-neutral MAB. It was shown[2] that the model-specific regret growth rate of MV-UCB was $O(\sqrt{T})$ and the model-independent regret growth rate of MV-DSEE was $O(T^{2/3})$.

Major questions that remain open are whether the $\sqrt{T}$ model-specific regret order and the $T^{2/3}$ model-independent regret order are the best one can hope for and whether the significant gaps in regret growth rate between risk-neutral MAB and risk-averse MAB (from $\log T$ to $\sqrt{T}$ in model-specific regret and from $\sqrt{T}$ to $T^{2/3}$ in model-independent regret) are inherent to the risk measure of mean-variance of observations. As shown in this work, the answer to these questions is negative in terms of model-specific regret and positive in terms of model-independent regret. Specifically, for model-specific regret, we establish an $\Omega(\log T)$ lower bound on the regret growth rate and provide a finer analysis of MV-UCB showing its $O(\log T)$ regret performance (in contrast to the $O(\sqrt{T})$ result given in [42]). In other words, the best achievable model-specific regret

[2]In [42], regret was defined comparing to the optimal single-arm policy that as we show in this work is not necessarily the optimal policy under a known model. However, we show that the difference between regret with regard to the optimal single-arm policy and the one with regard to the optimal policy is sufficiently small that preserves the order of the results (See Sec. 5.4). Also, in [42], a weaker regret definition, referred to as the pseudo regret, was considered. It was shown that the pseudo regret of MV-UCB was $O(\log^2(T))$. However, since the gap between pseudo regret and the strict regret is in the order of $O(\sqrt{T})$ (see Lemma 1 in [42]), the analysis in [42] only showed an $O(\sqrt{T})$ regret order of MV-UCB. We also point out that the two types of regret (model-specific vs. model-independent) were not distinguished in [42]. From their analysis, however, it is clear that the result on MV-UCB was in terms of model-specific regret while the result on MV-DSEE was in terms of model-independent regret.

		Risk-neutral MAB	Risk-averse MAB
Model-Specific	Lower Bounds	$\Omega(\log T)$ [1]	$\Omega(\log T)$
	Order-Optimal Policies	[1, 2, 3, 4]	MV-UCB
Model-Independent	Lower Bounds	$\Omega(\sqrt{T})$ [12, 11]	$\Omega(T^{2/3})$
	Order-Optimal Policies	[5]	MV-DSEE [42]

Table 5.1: Summary of results on risk-neutral and risk-averse MAB.

order remains to be logarithmic as in the risk-neutral MAB. In terms of model-independent regret, we show that the minimum regret growth rate is $\Omega(T^{2/3})$. Thus, the analysis of MV-DSEE given in [42] is tight. We thus complete in this work parallel results on risk-averse MAB under the measure of mean-variance of observations as summarized in the second column of Table 5.1.

5.1 Motivating Applications

Mean-variance is a well accepted risk measure whose quadratic scaling captures the natural inclination toward less risky options when the stakes are high. Studies have confirmed such risk-averse behaviors in investors (e.g. see [43]).

In the classic application of mean-variance to portfolio selection, the objective is a joint optimization of risk and return for a portfolio over a particular

period of time. This guarantees a high expected return and a low variation in the outcome. A similar approach is also taken for intertemporal returns of assets. Specifically, the objective is to guarantee high average return and low variations over time [45]. Such intertemporal variations are commonly referred to as *volatility* in finance literature and measured by the sample variance of the return process. The metric of mean-variance of the reward process studied in this work as well as in [42] and [46] precisely captures the objective of low volatility and high expected return. Another motivating application is clinical trial, where, besides obtaining high average return, it is desirable to avoid high variations in the treatment outcomes for different patients [42].

Another formulation of risk-averse MAB is to consider the mean-variance of the total return at the end of the time horizon where the objective is to minimize the ensemble variations of the total return. These two measures of mean-variance of the reward process and mean-variance of the total reward are suitable for different applications. For example, in the return of a financial security, the fluctuations over time are to be avoided as "risk for financial security" [47], while in a retirement investment, one might be more interested in the variation of the final return and less sensitive to the fluctuations in the intermediate returns. Some initial results on MAB under mean-variance of the total reward can be found in our preliminary study reported in [44].

5.2 Related Work

There are relatively few studies on risk-averse MAB. In an initial work on this topic, a sequential risk-averse problem using the measure of mean-variance of

observations was formulated in [46]. Different from this work and [42] that consider a stochastic formulation, [46] adopted the so-called non-stochastic full-information framework and established a negative result showing the infeasibility of sublinear regret.

There are a couple of results on risk-averse MAB under different risk measures. In [48], the quality of an arm was measured by a general function of the mean and the variance of the random variable. This study, however, is closer to the risk-neutral MAB problems than to the problem studied in this work. The reason is that under the model of [48], regret remains to be the sum of the immediate performance loss at each time instant. As discussed earlier, regret in mean-variance of observations is no longer summable over time.

In [44, 49], MAB under the measure of *value at risk*, which defines the minimum value of a random variable at a given confidence level, was studied. In [49], learning policies using the measure of conditional value at risk were developed. However, the performance guarantees were still within the risk-neutral MAB framework (in terms of the loss in the expected total reward) under the assumption that the best arm in terms of the mean value is also the best arm in terms of the conditional value at risk. In our recent work [44], we considered risk-averse MAB under the measure of value at risk of the total reward and developed learning policies that offer poly-log regret performance. Another risk measure for MAB problems was considered in [50] in which the logarithm of moment generating function was used as a risk measure and high probability bounds on regret were obtained.

There are also a couple of studies, while not directly addressing risk-averse MAB, offering relevant results from different perspectives. In [51], the sam-

ple complexity of both mean-variance and value at risk for single-period and multi-period decision making was studied. In [52], the problem of identifying the best arm in terms of different risk measures assuming the existence of an efficient risk estimator was considered. Identifying the best arm is, however, different from an MAB formulation due to the absence of the tradeoff between exploration and exploitation which is at the heart of online learning problems. Readers are also encouraged to read the work by Audibert *et al.* [53] on the deviation of regret from its expected value.

5.3 Notation and Preliminaries

Let $\{X_{\pi(t)}(t)\}_{t=1}^{T}$ denote the random reward sequence under policy π. The cumulative mean-variance $\xi_\pi(T)$ of the reward sequence is given in 5.2. The performance of policy π is measured by risk-averse regret $R_\pi(T)$ defined as the increase in cumulative mean-variance over a given horizon of length T as compared to the optimal policy π^* under a known model. (See Sec. 5.4 for a detailed discussion on π^*)

$$R_\pi(T) = \xi_\pi(T) - \xi_{\pi^*}(T). \tag{5.3}$$

5.3.1 Notations

Throughout the chapter, $*$ is used to indicate the arm that has the smallest mean-variance. If there are more than one arm with the smallest mean-variance value, one of them is chosen as $*$. Let $\Gamma_{i,j} = \mu_i - \mu_j$ and $\Delta_i = \xi_i - \xi_*$ denote, respectively, the difference between the mean values of arm i and j, and the difference between

the mean-variance of arm i and the arm with the smallest mean-variance. Let $\Delta = \min_{i \neq *} \Delta_i$, $\Gamma = \max_i |\Gamma_{i,*}|$, $\sigma_{\max} = \max_i \sigma_i$ and $\mu_{\max} = \max_i \mu_i$.

The following notations are used for the sample mean, the sample variance, and the sample mean-variance of the random reward sequence from arm i under a given policy π:

$$\begin{aligned} \overline{\mu}_i(t) &= \frac{1}{\tau_i(t)} \sum_{s=1}^{\tau_i(t)} X_i(t_i(s)), \\ \overline{\sigma^2}_i(t) &= \frac{1}{\tau_i(t)} \sum_{s=1}^{\tau_i(t)} (X_i(t_i(s)) - \overline{\mu}_i(t))^2, \\ \overline{\xi}_i(t) &= \overline{\sigma^2}_i(t) - \rho \overline{\mu}_i(t), \end{aligned}$$

where $t_i(s)$ denotes the time instant corresponding to the s'th observation from arm i and $\tau_i(t)$ denotes the number of times arm i has been played up to time t. Note that these quantities depend on the policy π, which is omitted for simplicity. The time argument may also be omitted when it is clear from the context. The use of the biased estimator for the variance is for the simplicity of the expression. The results presented in this work remain the same with the use of the unbiased estimator with $\tau_i(t)$ replaced by $\tau_i(t) - 1$ in the expression of $\overline{\sigma^2}_i(t)$.

The KL-divergence between two distributions f and g is given by

$$I(f, g) = \mathbb{E}_f[\log \frac{f(X)}{g(X)}], \tag{5.4}$$

where $\mathbb{E}_f$ denotes the expectation operator with respect to f.

In the proofs, the notation $\mathbb{E}[X, \mathcal{E}]$ for a random variable X and an event $\mathcal{E}$ is equivalent to $\mathbb{E}[X\mathbb{I}_{\mathcal{E}}]$, where $\mathbb{I}$ is the indicator function.

5.3.2 Concentration of the Sample Mean-Variance

We assume that $(X_i - \mu_i)^2 - \sigma_i^2$ $(i = 1, \ldots, K)$ for all arms have a sub-Gaussian distribution. Recall that a real-valued random variable X is called sub-Gaussian if it satisfies the following [29],

$$\mathbb{E}[e^{uX}] \leq e^{\zeta_0 u^2/2} \tag{5.5}$$

for some constant $\zeta_0 > 0$.

We establish in Lemma 1 a concentration result on the sample mean-variance, which plays an important role in regret analysis. This result is similar to the Chernoff-Hoeffding bound on the concentration of the sample mean for sub-Gaussian random variables [30]. The Chernoff-Hoeffding bound provides an upper bound on the probability of a given deviation of the sample mean from the true mean as given in 3.3. In the following lemma, we extend the Chernoff-Hoeffding bound to the sample mean-variance. Similar concentration inequalities for mean-variance were given in [42] and [51] for random variables with bounded support.

Lemma 2 *Let $\overline{\xi}_s$ be the sample mean-variance of a random variable X obtained from s i.i.d. observations. Let $\mu = \mathbb{E}[X]$, $\sigma^2 = \mathbb{E}[(X-\mu)^2]$, and assume that $(X-\mu)^2 - \sigma^2$ has a sub-Gaussian distribution, i.e.,*

$$\mathbb{E}e^{u(X-\mu)^2} \leq e^{\zeta_1 u^2/2}$$

for some constant $\zeta_1 > 0$. As a result $X - \mu$ has a sub-Gaussian distribution, i.e.,

$$\mathbb{E}[e^{uX}] \leq e^{\zeta_0 u^2/2}.$$

Let $\zeta = \max\{\zeta_0, \zeta_1\}$. We have, for all constants $a \in (0, \frac{1}{2\zeta}]$ and $\delta > 0$,

$$\mathbb{P}[\overline{\xi}_s - \xi(X) > \delta] \quad \leq 2\exp(-\tfrac{as\delta^2}{(1+\rho)^2}), \tag{5.6}$$

and for all $a \in (0, \frac{1}{2\zeta}]$ *and* $\delta \in (0, 2+\rho]$,

$$\mathbb{P}[\overline{\xi}_s - \xi(X) < -\delta] \quad \leq 2\exp(-\tfrac{as\delta^2}{(2+\rho)^2}). \tag{5.7}$$

Proof: See Appendix A.

5.4 The Known Model Case

In this section, we study the case where all arm distributions are known. This defines the benchmark performance in the regret definition given in 5.3. We first show through a counter example that playing the arm $*$ that has the smallest mean-variance may not be optimal. This presents a major difficulty in regret analysis given that explicit characterizations of the optimal policy π^* for the known model case are in general intractable. Our approach is to bound the performance gap between π^* and the optimal single-arm policy $\widehat{\pi}^*$ (i.e., playing arm $*$ all through), which allows us to analyze the order of the regret defined with respect to π^* by analyzing $\widehat{\pi}^*$.

To see that $\widehat{\pi}^*$ may not be optimal, the key is to notice that the variance term (i.e., the first term on the right-hand side of 5.2) in the cumulative mean-variance is with respect to the sample mean calculated from rewards obtained from all arms. When the remaining time horizon is short and the current sample mean is sufficiently close to the mean value of a suboptimal arm $j \neq *$, it may be more rewarding (in terms of minimizing the mean-variance) to play arm j rather than arm $*$. Consider a concrete example with two Gaussian-distributed arms with parameters $\mu_1 = 0$, $\mu_2 = 1$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2.1$. Let $\rho = 1$ and $T = 2$. It is easy to

see that $\xi_1 = 1$ and $\xi_2 = 1.1$, and the optimal single-arm policy $\widehat{\pi}^*$ is to always play arm 1, yielding a cumulative mean-variance of $\xi_{\widehat{\pi}^*}(t) = 1$. Consider a policy π with $\pi(1) = 1$ and $\pi(2) = \mathbb{I}_{X_1(1)<0.5} + 2\mathbb{I}_{X_1(1)\geq 0.5}$. It can be shown that $\xi_\pi(T) < 0.7$, demonstrating the sub-optimality of $\widehat{\pi}^*$.

The above example also gives a glimpse of the complexity in finding π^* for a general problem. To circumvent this difficulty, our approach is to show that $\widehat{\pi}^*$ is a good proxy of π^* with a performance loss upper bounded by a constant for large T. We can then obtain regret bounds through $\widehat{\pi}^*$.

Recall that regret $R_\pi(T)$ in 5.3 is defined with respect to π^*. Using $\widehat{\pi}^*$ as the benchmark, we define a proxy regret $\widehat{R}_\pi(T)$ as

$$\widehat{R}_\pi(T) = \xi_\pi(T) - \xi_{\widehat{\pi}^*}(T). \tag{5.8}$$

Our objective is to bound the difference between $R_\pi(T)$ and $\widehat{R}_\pi(T)$. To do this, we first derive in Lemma 3 a closed-form expression of $\widehat{R}_\pi(T)$ as a function of the number of times $\{\tau_i\}_{i=1}^K$ each arm is played over the entire horizon of length T. This lemma is the cornerstone of the regret analysis in subsequent sections.

Lemma 3 *The regret of a policy π with respect to the optimal single-arm policy $\widehat{\pi}^*$ under the measure of mean-variance of observations can be written as*

$$\widehat{R}_\pi(T) = \sum_{i=1}^K \mathbb{E}[\tau_i(T)]\Delta_i + \sum_{i=1}^K \mathbb{E}[\tau_i(T)]\Gamma_{i,*}^2 - \frac{1}{T}\mathbb{E}[(\sum_{i=1}^K \tau_i(T)(\overline{\mu}_i(T) - \mu_*))^2] + \sigma_*^2. \tag{5.9}$$

Proof: See Appendix A.

Recall that the regret in terms of the total expected reward can be written as a weighted sum of the expected value of $\tau_i(T)$. Specifically, based on Wald

identity, the regret is given by

$$\sum_{i=1}^{K} \mathbb{E}[\tau_i(T)](\mu_{\max} - \mu_i).$$

The regret in terms of mean-variance of observations is, however, a much more complex function of $\tau_i(T)$ as given in Lemma 3. It depends on not only the expected value of $\tau_i(T)$, but also the second moment of $\tau_i(T)$ and the cross correlation between $\tau_i(T)$ and $\tau_j(T)$.

Based on Lemma 3, we show in Theorem 12 that for $\Delta > 0$ and T sufficiently large, the difference between $R_\pi(T)$ and $\widehat{R}_\pi(T)$ is bounded by a constant independent of T.

Theorem 12 *For any policy π, we have*

$$0 \leq R_\pi(T) - \widehat{R}_\pi(T) \leq \min\{\sigma^2_{\max}(\sum_{i \neq *} \frac{\Gamma^2_{i,*}}{\Delta_i} + 1), \frac{K}{a} \log T\}. \tag{5.10}$$

Proof: Since the performance of the optimal policy cannot be worse than the optimal single-arm policy, we can immediately see that $\widehat{R}_\pi(T) \leq R_\pi(T)$. For the upper bound, we write $R_\pi(T) - \widehat{R}_\pi(T) = -\widehat{R}_{\pi^*}(T)$ and use the regret expression given in Lemma 3 to establish lower bounds on $\widehat{R}_{\pi^*}(T)$. We first show that for $\Delta > 0$ and large T, $\widehat{R}_{\pi^*}(T)$ is lower bounded by a constant. For the cases with small Δ, we show that, based on Lemma 4 (proved in Appendix A), the difference between the second and the third terms on the RHS of 5.9 is bounded by an order of $\log T$ term. For a detailed proof, see Appendix A.

Lemma 4 *Let $\{X(t)\}_{t=1}^T$ be an i.i.d. random process with mean $\mu = \mathbb{E}[X(t)]$ that satisfies 3.3 with constant a. Let $\tau \leq T$ be an stopping time for this random process and let $\overline{\mu}$ denote the sample mean from τ samples: $\overline{\mu} = \frac{\sum_{s=1}^{\tau} X(s)}{\tau}$. We have the following inequality*

$$\mathbb{E}[\tau(\overline{\mu} - \mu)^2] \quad \leq \quad \frac{1}{a}(\log T + 2). \tag{5.11}$$

5.5 Model-Specific Regret

In this section, we consider the model-specific setting. We establish lower bounds on model-specific regret feasible among all consistent policies and the order optimality of MV-UCB and MV-DSEE.

5.5.1 Lower Bounds on Model-Specific Regret

To avoid trivial lower bounds on regret caused by policies that heavily bias toward certain distribution models (e.g., a policy that always plays arm 1), the model-specific setting focuses on the so-called consistent policies. The model-specific lower bounds for risk-neutral MAB (Theorems 1 and 2 in [1]) are given for the set of policies that play suboptimal arms only $o(T^\alpha)$ times for all $\alpha \in (0,1)$. We relax this assumption and focus on $\alpha-$consistent policies defined as follows.

Definition. A policy π is α-consistent ($0 < \alpha < 1$) if for all reward distributions and for all $j \neq *$,

$$\mathbb{E}[\tau_j(T)] \leq T^\alpha. \tag{5.12}$$

We establish a lower bound on the model-specific regret feasible among the class of α-consistent policies for all $\alpha \in (0,1)$. Similar to the results by Lai and Robbins in [1] for risk-neutral MAB, we consider the family of one-parameter distribution models. Specifically, we assume that the distribution of arm i is

given by $f(.;\theta_i)$ and the distribution model $\mathcal{F} = (f(.;\theta_1), ..., f(.;\theta_K))$ can be represented by $\Theta = (\theta_1, ..., \theta_K)$. The parameters θ_i are taking value from a set $\mathcal{U}$ satisfying the following regularity condition (similar to that in [1]).

Assumption 1. For any θ, λ, and $\lambda' \in \mathcal{U}$, and for any $\epsilon > 0$, there exists a $\delta > 0$ such that $0 < \xi(\lambda') - \xi(\lambda) < \delta$ implies $|I(f(.;\theta), f(.;\lambda)) - I(f(.;\theta), f(.;\lambda'))| < \epsilon$.

The lower bound in [1] is asymptotic ($T \to \infty$). In addition to establishing the corresponding asymptotic lower bound for risk-averse MAB, we also provide in Theorem 13 a finite-time lower bound when the following assumption holds.

Assumption 2. For all θ and $\lambda \in \mathcal{U}$, let X be a sub-Gaussian random variable with distribution $f(.;\theta)$. The random variable $Y = f(X;\lambda)$ is sub-Gaussian[3].

Theorem 13 *Consider the MAB problem under the measure of mean-variance of observations. Let π be an $\alpha-$consistent policy and $\Theta \subset \mathcal{U}$ be the distribution model. Under Assumption 1, the model-specific regret satisfies, for any constant $c_1 < 1 - \alpha$, Furthermore, under Assumption 2, for $T_1 \in \mathbb{N}$, where ϵ_{T_1} can be arbitrary small when T_1 is large enough and $0 < c_2 < 1$ is independent of T and $\mathcal{F}$.*

proof 13 *The proof is based on the following lemma.*

Lemma 5 *Let Θ be the given distribution model, and let $i \neq *$ denote the index of a suboptimal arm under Θ. Let π be an $\alpha-$consistent policy. Under Assumption 1, the number $\tau_i(T)$ of times arm i is played under π satisfies, for any constant $c_1 < 1 - \alpha$,*

$$\lim_{T\to\infty} \mathbb{P}_{\mathcal{F}}[\tau_i(T) \geq \frac{c_1 \log T}{I(f_i, f_*)}] = 1, \tag{5.13}$$

[3]Note that Y is a function of X: for each $X = x$ generated according to $f(x;\theta)$, we have $Y = f(x;\lambda)$.

Furthermore, under Assumption 2, there exists $T_0 \in \mathbb{N}$ such that

$$\mathbb{P}_{\mathcal{F}}[\tau_i(T) \geq \frac{c_1 \log T}{I(f_i, f_*)}] \geq c_2, \quad \textit{for all } \ T > T_0, \tag{5.14}$$

where constant $0 < c_2 < 1$ is independent of T and $\mathcal{F}$.

To prove this lemma, we construct a new distribution model $\mathcal{F}^i$ where arm $i \neq *$ is the optimal arm. The log likelihood ratio γ between the two probability measures $\mathcal{F}$ and $\mathcal{F}^i$ is a key statistic to prove the lemma. Specifically, we show that it is unlikely that τ_i is smaller than the logarithmic term under two different cases of $\gamma \leq c_5 \log T$ and $\gamma > c_5 \log T$. The former is shown by a change of measure argument and using the consistency assumption. The latter is shown by Chernoff bound when Assumption 2 is satisfied and by law of large numbers otherwise. For a detailed proof see Appendix A.

To prove Theorem 13, we establish lower bounds on the first three terms of regret given in Lemma 3. Lemma 5 provides a lower bound on $\mathbb{E}[\tau_i]$. By showing a lower bound on the sum of the second and third terms we arrive at the theorem. For a detailed proof see Appendix A.

In comparison with Lai and Robbins lower bound for risk-neutral MAB [1], Theorem 13 considers a larger class of policies (by allowing a policy to be consistent with respect to a specific α rather than for all $\alpha \in (0, 1)$) and also provides a finite-time lower bound under Assumption 2. Note that the constant c_1 in Theorem 13 approaches one for policies that satisfy 5.12 for all $\alpha \in (0, 1)$, leading to a bound corresponding to that in [1].

5.5.2 Risk-Averse Learning Policies

The performance of MV-UCB was first analyzed in [42], which showed that the model-specific regret of MV-UCB was upper bounded by $O(\sqrt{T})$. Theorem 14 below gives a tighter analysis on the performance of MV-UCB, showing a $\log T$ regret order. This result, together with the lower bound given in Theorem 13, establishes the order optimality of the MV-UCB policy for the case of $\Delta > 0$.

MV-UCB assigns an index $\eta(t)$ to each arm and plays the arm with the smallest index at time t (after playing every arm once). The index depends on the sample mean-variance calculated from past observations and the number of times that the arm has been played up to time t. Specifically, the index of arm i at time t is given by

$$\eta_i(t) = \overline{\xi}_i(t) - b\sqrt{\frac{\log t}{\tau_i(t)}}, \tag{5.15}$$

where b is a policy parameter whose value depends on the risk measure (see Theorem 14 below).

Theorem 14 *Assume $\Delta > 0$. The regret offered by the MV-UCB policy with $b \geq \frac{\sqrt{3}(2+\rho)}{\sqrt{a}}$ under the measure of mean-variance of observations is upper bounded by*

$$\begin{aligned} R_{MV-UCB}(T) \quad \leq \quad & \sum_{\neq *}\left(\frac{4b^2 \log T}{\min\{\Delta_i^2, 4(2+\rho)^2\}} + 5\right)(\Delta_i + \Gamma_{i,*}^2) + \sigma_*^2 \\ & + \min\{\sigma_{\max}^2\left(\sum_{i \neq *} \frac{\Gamma_{i,*}^2}{\Delta_i} + 1\right), \frac{K}{a}\log T\}. \end{aligned} \tag{5.16}$$

proof 14 *From the regret expression given in 5.9, we need to first bound $\mathbb{E}[\tau_i]$ for $i \neq *$. This is established in the following lemma with proof given in Appendix A.*

Lemma 6 *Set $b \geq \frac{\sqrt{3}(2+\rho)}{\sqrt{a}}$. The expected number of times a sub-optimal arm $i \neq *$ with $\Delta_i > 0$ is played is upper bounded by*

$$\mathbb{E}[\tau_i(T)] \leq \frac{4b^2 \log T}{\min\{\Delta_i^2, 4(2+\rho)^2\}} + 5. \tag{5.17}$$

The third term in the regret expression in 5.9 is negative. Thus, we arrive at an upper bound on $\widehat{R}_{MV-UCB}(T)$ that translates to an upper bound on $R_{MV-UCB}(T)$ by applying Theorem 12. See Appendix A for a detailed proof.

The model-specific regret of MV-UCB is linear in T when $\Delta = 0$ as discussed in [42]. An alternative policy in this case is MV-DSEE, a variation of the DSEE policy developed in [4] for risk-neutral MAB. In the MV-DSEE policy, time is partitioned into two interleaving sequences: an exploration sequence denoted by $\mathcal{E}(t)$ and an exploitation sequence. In the former, the player plays all arms in a round-robin fashion. In the latter, the player plays the arm with the smallest sample mean-variance.

With the cardinality of the exploration sequence set to $\lceil f(T) \log T \rceil$ where $f(.)$ is a positive increasing diverging sequence with an arbitrarily slow rate, MV-DSEE offers an asymptotic regret order of $O(f(T) \log T)$ (which can be arbitrarily close to the optimal logarithmic order) over a fixed distribution model without the assumption of $\Delta > 0$.

Theorem 15 *The regret of MV-DSEE policy under the measure of mean-variance of observations is upper bounded by*

$$R_{MV-DSEE}(T) = O(f(T) \log T), \tag{5.18}$$

where $f(T)$ is a positive increasing diverging sequence with an arbitrarily slow rate.

proof 15 *Following similar steps as in the performance analysis of DSEE in the proof of Theorem 2, we can show that for $i \neq *$,*

$$\mathbb{E}[\tau_i] = O(f(T)\log T). \tag{5.19}$$

Also similar to the proof of Theorem 14, we have

$$R_{MV-DSEE}(T) \leq \sum_{i\neq *} \mathbb{E}[\tau_i(T)](\Delta_i + \Gamma^2_{i,*}) + \min\{\sigma^2_{\max}(\sum_{i\neq *} \frac{\Gamma^2_{i,*}}{\Delta_i} + 1), \frac{K}{a}\log T\}.$$

By substituting the bound on $\mathbb{E}[\tau_i]$ given in 5.19, we arrive at the theorem.

5.6 Model-Independent Regret

In this section, we consider the model-independent setting, in which the performance of a policy is measured against the worst-case reward model specific to the policy and the horizon length T. Specifically, let $R_\pi(T;\mathcal{F})$ denote the expected total performance loss of policy π over a horizon of length T for a reward model $\mathcal{F}$. The model-independent regret is given by, for each T,

$$R_\pi(T) = \sup_{\mathcal{F}} R_\pi(T;\mathcal{F}), \tag{5.20}$$

and we are interested in the order (in terms of T) of a thus defined $R_\pi(T)$. It is easy to see that for any MAB problem, the model-independent regret order cannot be lower than the model-specific regret order.

We establish an $\Omega(T^{2/3})$ lower bound on the model-independent regret of any policy. Specifically, in the following theorem we show that there is distribution model such that the regret grows with $\Omega(T^{2/3})$.

Theorem 16 *Consider the MAB problem under the measure of mean-variance of observations. The model-independent regret of any policy π satisfies, for some constants*

$c_3 > 0$ *and* $T_2 \in \mathbb{N}$,

$$R_\pi(T) \geq c_3 T^{2/3}, \quad \textit{for all } \ T > T_2. \tag{5.21}$$

proof 16 *The proof is based on a coupling argument between two bandit problems with* $K = 2$ *and under distribution models* $\mathcal{F}$ *and* $\mathcal{F}'$*, respectively. The optimal arm is switched between these two models while the difference* Δ *between the mean-variances of the optimal and the suboptimal arm is kept the same. First, it is shown that under at least one of these two distribution models, for some constants* $c_4 > 0$ *and* $T_2 \in \mathbb{N}$,

$$R_\pi(T) \geq \frac{c_4 \log T}{\Delta^2}, \quad \textit{for all } \ T > T_2. \tag{5.22}$$

Under both $\mathcal{F}$ *and* $\mathcal{F}'$*, a normal distribution is assigned to arm one. Two different Bernolli distributions are assigned to arm two such that arm two is the sub-optimal arm under* $\mathcal{F}$ *and the optimal arm under* $\mathcal{F}'$*. Through a coupling argument we show that for the specific distribution assignments designed here,*

$$\mathbb{P}_{\mathcal{F}}[\pi(t) = 2] + \mathbb{P}_{\mathcal{F}'}[\pi(t) = 1] \geq \exp(-\mathbb{E}_{\mathcal{F}}[\tau_2(T)] d_0 \Delta^2) \tag{5.23}$$

for some constant $d_0 > 0$*. A lower bound on regret can be derived from 5.23, which increases as* $\mathbb{E}_{\mathcal{F}}[\tau_2(T)]$ *decreases. On the other hand, a higher* $\mathbb{E}_{\mathcal{F}}[\tau_2(T)]$ *indicates a higher regret under distribution assignment* $\mathcal{F}$*. Optimizing the minimum of these two lower bounds for the value of* $\mathbb{E}_{\mathcal{F}}[\tau_2(T)]$ *leads to the desired lower bound in 5.22. A proper assignment of* $\Delta = d_6 T^{-\frac{1}{3}}$*, for some constant* d_6*, gives the lower bound on model-independent regret in 5.21. For a detailed proof, see Appendix A.*

MV-DSEE policy was also considered in [42] and was shown to achieve $O(T^{2/3})$ model-independent regret performance with the cardinality of the exploration sequence set to $|\mathcal{E}(T)| = \lceil T^{2/3} \rceil$. The lower bound given in Theorem 16 shows that MV-DSEE is order optimal under the model-independent setting.

5.7 Simulations

In this section, we provide numerical examples on the performance of MV-UCB. We first study the effect of risk tolerance factor ρ on the rewards obtained by a risk-averse policy. In Fig. 5.1, two sample returns of MV-UCB are shown (for $K = 4$, with normal reward distributions of parameters $\mu_1 = 0$, $\mu_2 = 1$, $\mu_3 = 2$, $\mu_4 = 3$, $\sigma_1 = 1$, $\sigma_2 = 1$, $\sigma_3 = 2$, $\sigma_4 = 2$). By decreasing ρ the variation in the observation decreases, although it is at a price of a lower average return.

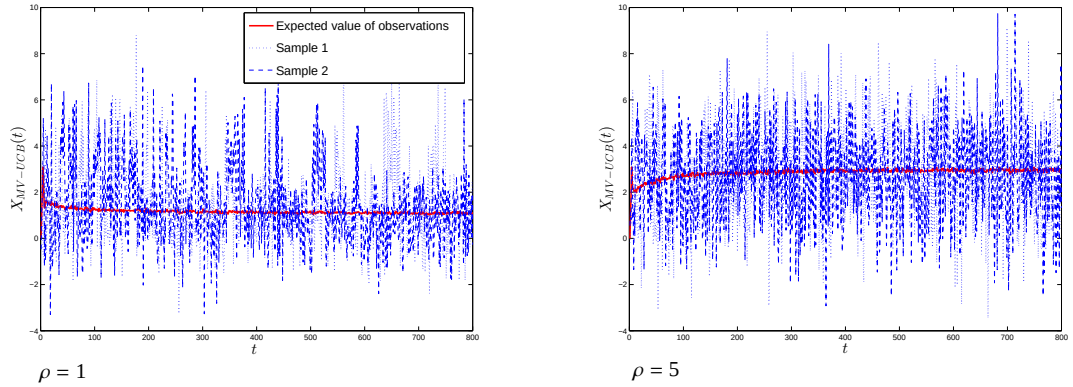

Figure 5.1: The sample observations of MV-UCB under different risk-tolerance factor ρ.

Fig. 5.2 shows the regret performance of MV-UCB for different values of Δ. The simulation shows that for a fixed value of Γ, the regret offered by MV-UCB increases as Δ decreases. A linear regret order is expected as Δ approaches 0.

5.8 Discussion

We studied risk-averse MAB problems under the risk measure of mean-variance of observations. We fully characterized the regret growth rate in both the model-

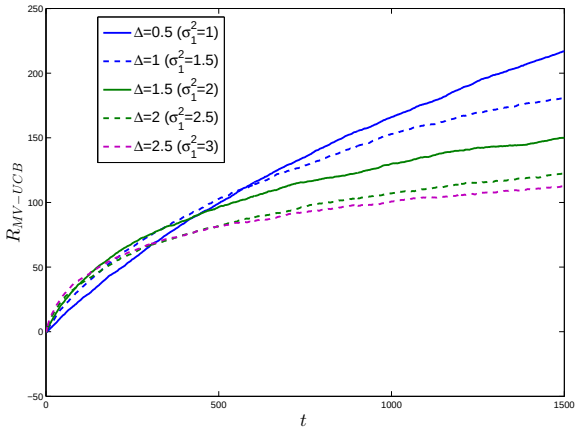

Figure 5.2: The performance of MV-UCB ($\rho = 1$, $K = 2$ with normal reward distributions of parameters $\mu_1 = 0$, $\mu_2 = 0.5$, $\sigma_2^2 = 1$).

specific and the model-independent settings by establishing lower bounds and developing order-optimal online learning policies.

The risk-averse MAB model reduces to the classic risk-neutral MAB when $\rho \to \infty$. Specifically, when $\rho \to \infty$, the mean-variance approaches to the negative of the mean multiplied by ρ. Thus, the mean-variance measure degenerates to a scaled mean value measure. With Δ_i replaced by $-\rho\Gamma_{i,*}$ and $\Gamma_{i,*}^2$ negligible against the term $-\rho\Gamma_{i,*}$, the model-specific bounds given in Theorems 13 and 14 reproduce the bounds on risk-neutral regret. Regarding the model-independent regret, however, as it is shown in this work, the regret growth rate is different from the risk-neutral MAB. This difference is expected due to the reason that the worst-case assignment of the distributions takes into account the value of ρ. Thus, even for a large value of ρ, a proper choice of the distributions with a sufficiently small difference $\Gamma_{i,*}$ in the mean values results in a case where the difference in variance is comparable with $-\rho\Gamma_{i,*}$ and cannot be ignored.

The model-specific regret lower bound obtained in Theorem 13 applies to

only single-parameter distribution models, the same as the lower bound obtained by Lai and Robbins in [1] for risk-neutral MAB. Under the measure of mean-variance of observations, the mean and the mean-variance of each arm are dependent through the single parameter θ_i of the distribution. Thus, the values of Δ_i and $\Gamma_{i,*}$ cannot be set independently. As a result, the $\Omega(T^{\frac{2}{3}})$ regret lower bound in the model-independent setting cannot be deduced from Theorem 13. The regret performance of MV-UCB and MV-DSEE as given in Theorem 3 and Theorem 4, however, does not require the assumption of single-parameter distribution models. It is thus perhaps reasonable to expect that the logarithmic order in the lower bound holds for general distribution models.

The time variations in the reward process have two sources: the randomness of the observed reward from each arm and the switching across arms with different expected values. The latter diminishes when $\Gamma \to 0$. Consequently, when $\Gamma \to 0$, the regret in mean-variance of observations becomes summable over time and is given by a weighted sum of the expected number of times that each suboptimal arm is played with the weights given by the difference in the variance of a suboptimal arm from the optimal arm. It is thus similar to the risk-neutral regret with the difference in mean replaced by the difference in variance. Thus, as expected, the model-specific bounds given in Theorems 13 and 14 degenerate to the bounds on risk-neutral regret, except that Δ_i is the difference in the variance rather than the mean. Under the model-independent setting, the value of Γ is chosen for the worst-case assignment of the distribution model and cannot be forced to zero. The above connection through $\Gamma \to 0$ between the regret in mean-variance and the regret in mean is thus absent in the model-independent setting.

Our regret lower bounds that hold for all $T \geq T_0$ for some constant $T_0 \in \mathbb{N}$ should be interpreted as finite-time results since one can always find a leading constant large enough (in the case of upper bounds) or small enough (in the case of lower bounds) to accommodate the first T_0 terms. Indeed, how large or small the leading constant needs to be to have results hold for all T can be obtained in our proof procedure. However, such a practice is tedious and leads to an overly complicated expression.

For the risk-neutral MAB, an improved version of the UCB policy developed in [5] was shown to achieve the optimal regret order under both the model-specific and model-independent settings. We have shown in this work that MV-DSEE approaches both the model-specific and model-independent regret lower bounds, but requiring different values for the cardinality of the exploration sequence. Whether a single policy without any change in its parameter values can achieve the optimal regret order under both settings remains an open question. A satisfactory answer to this question is involved and requires a separate investigation.

CHAPTER 6

RISK-AVERSE MULTI-ARMED BANDIT UNDER VALUE AT RISK MEASURE

The classic MAB formulation targets at maximizing the *expected* return of an online learning policy. In many applications, especially in economics and finance, a player may be more interested in reducing the uncertainty (i.e., risk) in the outcome, rather than achieving the highest ensemble average. In this chapter we develop a learning policy for risk-averse MAB under the measure of Value at Risk (VaR). Commonly used in economics and financial mathematics, for a random variable X and a given probability $p \in [0, 1]$, the Value at Risk (VaR) $\nu_p(X)$ is defined as

$$\nu_p(X) = \inf\{x \in \mathbb{R} : \Pr[X \leq x] > p\}. \tag{6.1}$$

VaR can be considered as the quantile or the inverse CDF of X. It bounds the minimum value of X (e.g., the minimum return of an investment) at a *confidence level* of $1 - p$.

We aim at designing learning policies to maximize the value at risk (potentially a negative value) in the total reward. More specifically, let $\{Z_t\}_{t=1}^T$ denote the sequence of rewards obtained at times $t = 1, ..., T$. The total reward $W^{(T)} = \sum_{t=1}^T Z_t$ is a random variable where the randomness comes from the randomness in the rewards and possible randomness in the arm selection policy. A risk-averse MAB problem under the VaR measure will be interested in a policy that maximizes $\nu_\epsilon(W^{(T)})$ where ϵ is an arbitrarily small probability determining the confidence level $1 - \epsilon$.

To give intuition into this risk-averse MAB problem, in Sec. 6.2, we focus on a K-armed bandit with Gaussian distrusting of the rewards. We show that, although the VaR is a non-linear measure of the random variables, the optimal policy is still to play a single best arm. We also show that, with a non zero-gap in the mean values of the arms the optimal policy, asymptotically, is still to play the arm with the largest mean. This argument also clarifies that the classic online learning policies such as UCB designed for risk-neutral MAB [3] achieve the optimal order of the risk-averse regret. However, an important question that remains to be answered is that in a model where the arms are equivalent in terms of their mean, what would be the second measure to determine the quality of arms in a risk-averse sense. We show that the answer to this question, as it might sound intuitive, is to play the arm with the lowest variance (in case of Gaussian distributions). We introduce a policy, similar to UCB, that uses the sample variance to obtain a lower confidence bound (LCB) on the variance of each arm and plays the arm with the smallest LCB. We show that this policy achieves the optimal logarithmic number of playing suboptimal arms following the similar lines as in the analysis of UCB as presented in [3].

In Sec. 6.3, we study the general case where the distribution of rewards are not necessarily Gaussian but sub-Gaussian. We show that the similar results as in the Gaussian case hold. Except, the second determining measure (besides the expected value of the rewards) is a normalized log moment generating function (N-log-MGF) instead of variance. Designing a learning policy that plays the arm with the smallest N-log-MGF is a more involved problem because it is no longer feasible to use a sample mean estimation to obtain a confidence bound on N-log-MGF. We introduce a policy (inspired by KL-UCB [65] and RA-UCB [50]) that uses the empirical distribution obtained from the past observation to play

the best arm. We show that the proposed policy, referred to as VaR-Learning (VaR-L), obtains the optimal order logarithmic number of playing the suboptimal arms.

6.1 Notation and Preliminaries

Let us use the notation $X_{k,s}$ for the s's observation from arm k. For the simplicity of presentation of the proofs we use the following concise notations: $Z_t = X_{\pi(t)}(t)$ and $W^{(T)} = \sum_{t=1}^{T} X_{\pi(t)}(t)$ wherever π is clear from context.

In risk-averse MAB problem under VaR measure, the objective is to design policies that return higher VaR. The performance of a policy π, parallel to the classic MAB, is measured by regret defined as the loss in the VaR over a given horizon of length T as compared to the optimal policy π^* under known model.

$$R_{\epsilon}^{(\pi)}(T) = \nu_{\epsilon}(\sum_{t=1}^{T} X_{\pi^*(t)}(t)) - \nu_{\epsilon}(\sum_{t=1}^{T} X_{\pi(t)}(t)). \tag{6.2}$$

For a random variable X, let $\mu(X)$ and $\sigma^2(X)$ denote the mean and variance of X, respectively.

The KL-divergence between two distributions f and g is given by

$$\mathcal{D}(f, g) = \mathbb{E}_f[\log \frac{f(X)}{g(X)}], \tag{6.3}$$

where $\mathbb{E}_f$ denotes the expectation operator with respect to f.

For a random variable X with distribution f, the N-log-MGF at a parameter λ is denoted by $\eta_\lambda(f)$ and defined as

$$\eta_\lambda(f) = \frac{1}{\lambda^2} \log \mathbb{E}_f[\exp(\lambda X)]. \tag{6.4}$$

The notation $\eta_{i,\lambda} = \eta_\lambda(f_i)$ is also used for the N-log-MGF of the reward of arm i. The error function denoted by ϕ is defined as

$$\phi(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-z^2} dz. \tag{6.5}$$

6.2 Gaussian Reward Distribution Model

To start with we consider a K-armed bandit with all Gaussian distributions. We show that the classic learning policies such as UCB designed for risk-neutral MAB achieve the optimal order regret in a risk averse MAB formulation where there is a gap in the mean values of the arms. However, in a model where the arms are equivalent in terms of their mean a policy that plays the arm with the smallest variance archives the highest VaR. We introduce σ-LCB policy that obtains the optimal order of the risk-averse regret.

By definition of error function, we can see that if X is a Gaussian random variable with mean $\mu(X)$ and variance $\sigma^2(X)$, for the VaR of X at a confidence level $1 - \epsilon$, we have

$$\nu_p(X) = \mu(X) + \sqrt{2\sigma^2(X)}\phi^{-1}(2\epsilon - 1) \tag{6.6}$$

Based on property 6.6, we give an expression for the VaR of the total reward of a policy π in the following theorem.

Theorem 17 *In a MAB problem under VaR measure where all distributions in $\mathcal{F}$ are Gaussian we have the following expression for the VaR of the total reward of an arm selection policy π*

$$\nu_\epsilon(\sum_{t=1}^{T} X_{\pi(t)}(t)) = \sum_k \mathbb{E}[\tau_k]\mu_k + \sqrt{2}\phi^{-1}(2\epsilon-1)\sqrt{\sum_k \mathbb{E}[\tau_k]\sigma_k^2}. \quad (6.7)$$

proof: The sum of independent Gaussian random variables is a Gaussian random variable. Equation 6.7 can be shown based on 6.6 and the mean and the variance of $\sum_{t=1}^{T} X_{\pi(t)}(t)$. ■

Using standard inequalities for error function we can show that (notice that $\phi^{-1}(2\epsilon-1) < 0$)

$$c_1\sqrt{\log\frac{c_2}{\epsilon}} \le -\phi^{-1}(2\epsilon-1) \le c_3\sqrt{\log\frac{c_4}{\epsilon}}, \quad (6.8)$$

for some constants $\frac{1}{\sqrt{2}} \le c_1, c_2, c_3, c_4 \le 1$. Thus, the second term in 6.7 scales in order of $\sqrt{\log\frac{1}{\epsilon}}$ with ϵ as ϵ goes to zero.

We prove that a policy π^*, playing one single arm with the largest $\mu_k + \phi^{-1}(2\epsilon-1)\sqrt{\frac{2\sigma_k^2}{T}}$, achieves the highest VaR. We refer to this policy as the optimal policy under known model and use * as the index of the single best arm (μ_* and σ_*^2 denote the mean and variance of the single best arm).

Lemma 7 *The VaR of a policy π defined in 6.7 is maximized when only a single arm with the largest $\mu_k + \phi^{-1}(2\epsilon-1)\sqrt{\frac{2\sigma_k^2}{T}}$ is played.*

proof: Define

$$\nu_\epsilon(t_1, t_2, ...t_k) = \sum_k t_k\mu_k + \sqrt{2}\phi^{(-1)}(2\epsilon-1)\sqrt{\sum_k t_k\sigma_k^2}. \quad (6.9)$$

over real numbers $(t_1, t_2, ...t_k) \in [0, T]^K$ with $K-1$ degree of freedom such that $t_* = T - \sum_{k \neq *} t_k$. We then take the second derivative of $v_\epsilon(t_1, t_2, ...t_k)$ with respect to t_k for $k \neq *$ and show that $\frac{\partial^2 v_\epsilon}{\partial t_k^2}$ is positive over $0 \leq t_k \leq T$. Thus, v_ϵ is maximized at one of the two ends: $t_k = 0$ or $t_k = T$. Then, we check that $\tau_* = T$ maximizes v_ϵ which implies the same result for $v_\epsilon(\sum_{t=1}^{T} X_{\pi(t)}(t))$. ■

From Lemma 7, we can see that playing the single best arm $*$ results in the highest VaR. Thus, we can rewrite the expression of the risk-averse regret as

$$R_\epsilon^{(\pi)}(T) = \sum_k \mathbb{E}[\tau_k] \Big(\mu_* - \mu_k + \sqrt{2}\phi^{-1}(2\epsilon - 1) \frac{\sigma_*^2 - \sigma_k^2}{\sqrt{T\sigma_*^2} + \sqrt{\sum_k \mathbb{E}[\tau_k]\sigma_k^2}} \Big). \tag{6.10}$$

As we can see from the above expression, the second term inside the brackets in 6.10 decreases with T in an $O(\frac{1}{\sqrt{T}})$. From here, we can conclude that the risk-neutral MAB policies such as UCB would work optimal up to constants or in order in a risk-averse MAB problem unless the arms are identical in terms of their mean. We also point out that the risk-averse regret degenerates to the classic regret asymptotically as T grows large because the the second term inside the brackets in 6.10 goes to zero as fast as $O(\frac{1}{\sqrt{T}})$. Thus, both the risk-neutral and the risk-averse regrets show the same asymptotic behavior when there is a gap in the mean values.

Classic MAB policies generally focus on the mean value of the arms and use the sample mean of the past observations as an estimate of mean to choose the best arm to play. An interesting question that arises here is that which arms are preferred to be played when the arms have the same mean values. For a MAB problem under VaR measure with Gaussian distributions the answer to

this question is to play the arm with the lowest variance.

We formalize this intuitive conclusion in the rest of this section and introduce σ-LCB policy that achieves the order optimal performance.

Consider a K-armed bandit with Gaussian distributions with the same mean $\mu_k = \mu_*$ for all $k = 1, ..., K$. Clearly, the regret in the expectation is zero for any arm selection policy π. However, the risk-averse regret defined in 6.2 is still an appropriate measure to compare learning policies.

The policy σ-LCB, similar to UCB policy assigns an index $I(t)$ to each arm and plays the arm with the smallest index at time t (after playing each arm once). The index depends on the sample variance calculated from the past observations and the number of times each arm has been played up to time t. Specifically, the index of arm k is given by

$$I_k(t) = \overline{\sigma}_k^2 - b\sqrt{\frac{\log t}{\tau_k(t)}}, \tag{6.11}$$

for some constant b. The value of b can be set to any constant such that $b \geq 4\sqrt{6\sigma_{(U)}^2}$ where $\sigma_{(U)}^2$ is an upper bound on the variances of the arms.

Theorem 18 *In a K-armed bandit problem with $\mu_k = \mu_*$ for all $k = 1, ..., K$, the regret offered by σ-LCB policy is upper bounded by*

$$R_\epsilon^{(\pi)}(T) \leq \sqrt{2}\phi^{-1}(2\epsilon - 1)\frac{\sum_k u_k(\sigma_*^2 - \sigma_k^2))}{\sqrt{T\sigma_*^2} + \sqrt{\sum_k u_k\sigma_k^2}} \tag{6.12}$$

Where u_k are upper bounds on $\tau_k(T)$ for all $k \neq *$ and $u_* = T - \sum_{k\neq *} u_k$. For $k \neq *$, u_k are given as

$$u_k = \frac{4b^2 \log T}{\sigma_i^2 - \sigma_*^2} + 5. \tag{6.13}$$

proof: First, we show that for a normal distribution, the following concentration inequalities hold on the sample variance obtained from s observations $\{X_1, X_2, ..., X_s\}$

$$\mathbb{P}[\frac{1}{s}\sum_{i=1}^{s}(X_i - \frac{1}{s}\sum_{i=1}^{s}X_i)^2 - 1 < -\delta] \leq 2e^{-s\delta^2/32}$$
$$\mathbb{P}[\frac{1}{s}\sum_{i=1}^{s}(X_i - \frac{1}{s}\sum_{i=1}^{s}X_i)^2 - 1 > \delta] \leq e^{-s\delta^2/8}$$

We then follow the similar lines as in the analysis of UCB [3] or MV-UCB [**?**] to prove an upper bound on $\mathbb{E}[\tau_k(T)]$. ■

6.3 Sub-Gaussian Reward Distribution Model

With general reward distributions, the VaR of a random variable is a non-linear function of the distributions. The exact VaR of the accumulative reward is intractable in general. The VaR for a general model can no longer be written as a linear combination of mean and standard deviation. However, it is still possible to approximate the VaR with a linear combination of mean and standard deviation up to a constant that scales with $\frac{1}{\epsilon}$ (see [44] which uses Berry-Esseen theorem for such approximations). Given the second term in regret 6.10 scales with $\sqrt{\log\frac{1}{\epsilon}}$ this approximation is a rather loose and inefficient approximation for small values of ϵ. In this paper we introduce a finer approach to VaR with general distributions building on the following concentration bounds.

Lemma 8 *For a sub-Gaussian random variable X let $\mu = \mathbb{E}[X]$ and $\overline{X}_s$ denote the sample mean obtained from s observations. We have, for some $\lambda_0 > 0$,*

$$\mathbb{P}\Big[\overline{X}_s - \mu > \frac{2}{\lambda_0}\sqrt{\frac{1}{s}\log\frac{1}{\epsilon}\log M(\lambda_0)}\Big] \leq \epsilon$$
$$\mathbb{P}\Big[\overline{X}_s - \mu < \frac{-2}{\lambda_0}\sqrt{\frac{1}{s}\log\frac{1}{\epsilon}\log M(-\lambda_0)}\Big] \leq \epsilon. \tag{6.14}$$

proof:

We multiply both sides of the inequality inside the probability measure with λ_0 and then apply exponential function to both sides. Using Markov inequality and optimizing on the value of λ_0 we arrive at 6.14. ■

Based on the above concentration inequalities, parallel to Gaussian model, we introduce the following expression for the risk-averse pseudo-regret.

$$\widetilde{R}_\epsilon^{(\pi)}(T) = \sum_k \mathbb{E}[\tau_k](\mu_* - \mu_k)$$
$$+2\sqrt{\log\frac{1}{\epsilon}}\Big(\sqrt{T\eta_{*,\lambda_0}} - \sqrt{\sum_k \mathbb{E}[\tau_k]\eta_{k,\lambda_0}}\Big). \tag{6.15}$$

The above expression is called pseudo-regret in the sense that instead of optimizing the VaR an upper bound on the VaR (that can be derived from the concentration inequalities 6.14) is optimized. Notice that, for a Gaussian distribution, the N-log-MGF equals to variance divided by 2. Thus, the pseudo-regret given in 6.15 consistently degenerates to the regret given in 6.10 for Gaussian distributions up to approximation of $\phi^{-1}(2\epsilon - 1)$.

Following the similar lines of argument as in Sec. 6.2, we can see that the second term in regret expression in 6.15 is smaller than the first term in order. Thus the classic MAB policies achieve optimal order performance. However, the interesting question remains that when the arms are equivalent in terms of their mean which arms are preferred to be played. From regret expression in 6.15,

the answer to this question is to play the arms with the smallest N-log-MGF at a parameter λ_0.

We introduce the VaR-L policy in this section that achieves the optimal logarithmic order of of playing suboptimal arms. Let $\widehat{f_{k,s}}$ be the empirical distribution of the arm k based on past s observations.

$$\widehat{f_{k,s}} = \frac{1}{s}\sum_{i=1}^{s}\delta_{X_k,i}. \tag{6.16}$$

The VaR-L policy assigns an index $J(t)$ to each arm and plays the arm with the smallest index.

$$\pi(t+1) = \arg\min_k J_k(t). \tag{6.17}$$

The index $J_k(t)$ for each arm is defined as

$$J_k(t) = \inf\left\{\alpha : d(\widehat{f_{k,\tau_k(t)}}, \alpha) \le \frac{a(t)}{\tau_k(t)}\right\}, \tag{6.18}$$

where

$$d(\widehat{f_i}, \alpha) = \inf_f\{\mathcal{D}(\widehat{f_{i,\tau_i(t)}}, f) : \eta_\lambda(f) \le \alpha\}, \tag{6.19}$$

and $a(t)$ is an increasing sequence in time with $a(t) = O(\log T)$. It has been shown in [50] that the above optimization problem that determines the index $J(t)$ is efficiently calculable.

In Theorem 19, we establish an upper bound on the number of times each suboptimal arm is played by VaR-L.

Theorem 19 *In a K-armed bandit, the number of times each suboptimal arm $k \neq *$ is played by VaR-L policy satisfies*

$$\mathbb{E}[\tau_k(T)] = O(\log T). \tag{6.20}$$

proof:

For an arm $k \neq *$, in order to have $\pi(t) = k$, at least one of the two following events should be satisfied: $\{J_*(t) \geq \eta_{*,\lambda}\}$ or $\{J_k(t) < \eta_{*,\lambda} \ , \ \pi(t) = k\}$. The former implies that $d(\widehat{f}_{*,\tau_*(t)}, \eta_{*,\lambda}) \geq \frac{a(t)}{\tau_*(t)}$. The latter implies that $d(\widehat{f}_{k,\tau_k(t)}, \eta_{*,\lambda}) \leq \frac{a(t)}{\tau_k(t)}$. We then follow the similar lines as in the analysis of RA-UCB [50] where upper bounds on the probabilities of $\mathbb{P}[d(\widehat{f}_{*,\tau_*(t)}, \eta_{*,\lambda}) \geq \delta]$ and $\mathbb{P}[d(\widehat{f}_{k,\tau_k(t)}, \eta_{*,\lambda}) \leq \delta]$ were established. By using the upper bounds for $\delta = \frac{a(t)}{\tau(t)}$ we arrive at an upper bound on $\mathbb{E}[\tau_k(T)]$. The analysis of RA-UCB is slightly different in the sense that the normalization of log-MGF is different. ■

6.4 Conclusion

We showed that the risk-neutral learning policies such as UCB achieve the optimal order regret in a risk averse MAB under VaR measure where there is a gap in the mean values of the arms. In a model where the arms are equivalent in terms of their mean we introduced a learning policy that achieves the highest order VaR by playing suboptimal arms in a logarithmic order with time.

CHAPTER 7

ACHIEVING BOUNDED REGRET WITH MINIMAL SIDE INFORMATION

The minimum logarithmic regret growth rate in MAB indicates that mistakes in selecting a suboptimal arm occur infinitely often, and any uniformly good online learning policy never converge to the best arm (i.e., the arm with the largest reward mean). It is thus interesting to ask whether any side information on the reward model can lead to bounded regret, thus complete learning, and if yes, what is the minimum side information for achieving complete learning. An initial attempt at these questions was made in [11] where it was shown that when the value of the largest reward mean (among all arms) and a positive lower bound on the difference between the largest and the second largest reward mean is known, the regret is bounded over time. In other words given the mean values of the best arm and an upper bound on the mean value of the second best arm, the regret is bounded rather than growing logarithmically with time as shown by Lai and Robbins in [1]. It was also shown in [11] that if only one of these two pieces of information is known, the regret is still logarithmic with time.

One may then wonder whether what was shown in [11] is a set of minimum side information for achieving bounded regret. We show in this work that the answer is negative. Specifically, we show that the knowledge of a value η between the largest and the second largest reward mean is sufficient to achieve bounded regret. It is easy to see that the knowledge of such an η is a smaller piece of side information than what was required in [11] because the latter leads to the former but not vice verse. Furthermore, our result applies to both

light-tailed and heavy-tailed reward distributions.

Consider an ordering between two pieces of information $\mathcal{I}$ and $\mathcal{I}'$. We say $\mathcal{I}$ is strictly less than $\mathcal{I}'$ denoted by $\mathcal{I} \prec \mathcal{I}'$ if given $\mathcal{I}'$, $\mathcal{I}$ is known but not vice versa. We then say that $\mathcal{I}_{min}$ is a set of minimum side information for achieving bounded regret if

- there exists a policy that achieves a bounded regret with the knowledge of $\mathcal{I}_{min}$;
- for any $\mathcal{I}' \prec \mathcal{I}_{min}$, no policy can achieve bounded regret with the knowledge of $\mathcal{I}'$.

7.1 Bounded Regret Policy

In this section we propose a learning policy that achieves bounded regret provided a value η between the largest and the second largest reward mean. We provide analysis for the regret performance of the proposed policy under Sub-Gaussian and heavy-tailed distribution models.

7.1.1 Under Sub-Gaussian Distribution Model

Let σ be a permutation of $\{1, ..., K\}$ such that $\mu_{\sigma(1)} \geq \mu_{\sigma(2)} \geq \cdots \geq \mu_{\sigma(K)}$ and Δ_k^σ be the gap in the mean values of arm $\sigma(k)$ and arm $\sigma(1)$ (i.e. $*$). We might drop the superscript σ from Δ_k^σ when it is clear from the context. Given a value η with $\mu_{\sigma(2)} \leq \eta < \mu_{\sigma(1)}$ our proposed policy π_η which achieves bounded regret for Sub-Gaussian distributions is as follows. Initially play each arm once in a round

robin fashion. Then find the set $\mathcal{S}_t$ defined as

$$\mathcal{S}_t = \{k : \bar{\mu}_k(t) > \eta + \tau_k(t)^{-1/3}\}. \tag{7.1}$$

If $\mathcal{S}_t \neq \emptyset$ play the arm in $\mathcal{S}_t$ with the largest $\bar{\mu}_k(t)$. If $\mathcal{S}_t = \emptyset$, in a round robin fashion, play each arm one time in the next K rounds. The algorithm is described in Fig. 7.1.1.

π_η Policy

- Notation: Define the set $\mathcal{S}_t = \{k : \bar{\mu}_k(t) > \eta + \tau_k(t)^{-1/3}\}$ at each time t.

1. play each arm once in a round robin fashion
2. *while* $\mathcal{S}_t \neq \emptyset$
   - play the arm with the largest $\bar{\mu}_k(t)$ in $\mathcal{S}_t$
3. go to line 1

Figure 7.1: The description of the π_η policy.

Theorem 20 *Under a Sub-Gaussian distribution model, the regret of the proposed policy π_η is bounded by a constant for all T. Specifically for any $s_0 \in \mathbb{N}$ such that $\delta_0 \triangleq \mu_{\sigma(1)} - \eta - s_0^{-1/3} > 0$,*

$$R_{\pi_\eta}(T) \leq \frac{6}{a^3}(\mu_{\sigma(1)} - \eta)(K-1) + \sum_{k=2}^{K}[s_0 + \frac{1}{a\delta_0^2}](\mu_{\sigma(1)} - \mu_{\sigma(k)}),\ \forall\ T.$$

We point out that in the policy π_η we can substitute $-1/3$ in $\tau_k(t)^{-1/3}$ with any real number $\kappa \in (-1/2, 0)$. Following exactly the same lines of the proof of the Theorem 20, we can show that for any $\kappa \in (-1/2, 0)$ the regret is bounded by,

$$\begin{aligned} R_{\pi_\eta}(T) &\leq \int_0^\infty \exp(-au^{1-2\kappa})du(\mu_{\sigma(1)} - \eta)(K-1) \\ &+ \sum_{k=2}^{K}[s_0 + \frac{1}{a\delta_0^2}](\mu_{\sigma(1)} - \mu_{\sigma(k)}), \end{aligned} \tag{7.2}$$

where $\delta_0 \triangleq \mu_{\sigma(1)} - \eta - s_0^{\kappa} > 0$.

Let δ denote the lower bound on the difference between the largest and the second largest reward mean. The side information for our proposed policy is the value η. From the definition of $\prec$ the following holds:

- $\eta \prec (\mu_{\sigma(1)}, \delta)$

Thus, our proposed policy needs strictly less side information to achieve complete learning comparing with [11] which needs (θ_1, δ) as side information.

7.1.2 Under Heavy-Tailed Distribution Model

If we substitute the mean estimator $\bar{\mu}_k(t)$ in π_η with a carefully chosen truncated sample mean, the bounded regret is achievable for all heavy-tailed distributions with pth moment ($p > 1$).

Assume for $p \in (1, 2]$ and some $u < \infty$,

$$\mathbb{E}[|X_k(t)|^p] \leq u. \tag{7.3}$$

Let $a = 4^{p/1-p} u^{1/1-p}$, and

$$\epsilon_s = \exp(-a s^{1+\kappa p/p-1}), \tag{7.4}$$

for some $\kappa \in (\frac{1-p}{p}, 0)$.

Given an η with $\mu_{\sigma(2)} \leq \eta < \mu_{\sigma(1)}$, let policy $\widetilde{\pi}_\eta$ be as follows. Define the truncated empirical mean $\tilde{\mu}_k(t)$ as

$$\tilde{\mu}_k(t) = \frac{1}{\tau_k(t)} \sum_{s=1}^{\tau_k(t)} X_k(s) \mathbb{I}\left\{ |X_k(s)| \leq \left(\frac{us}{\log(\epsilon_s^{-1})} \right)^{1/p} \right\}. \tag{7.5}$$

Initially play each arm once in a round robin fashion. Then find the set $\mathcal{S}_t$ defined as $\tilde{\mathcal{S}}_t = \{k : \tilde{\mu}_k(t) > \eta + \tau_k(t)^\kappa\}$. If $\tilde{\mathcal{S}}_t \neq \emptyset$ play the arm in $\tilde{\mathcal{S}}_t$ with the largest $\tilde{\mu}_k(t)$. If $\tilde{\mathcal{S}}_t = \emptyset$, in a round robin fashion, play each arm one time in the next K rounds.

The following theorem states the regret performance of the policy $\tilde{\pi}_\eta$.

Theorem 21 *Assume that the reward distribution for each arm k satisfies* $\mathbb{E}[|X_k|^p] \leq u$ *for some* $p \in (1, 2]$, *and* $u > 0$. *Let* $a = 4^{p/p-1}u^{1/p-1}$. *Then for any* $\kappa \in (\frac{1-p}{p}, 0)$, *the regret of the proposed policy* $\tilde{\pi}_\eta$ *is bounded by a constant for all* T. *Specifically for any* $s_0 \in \mathbb{N}$ *such that* $\delta_0 \triangleq \mu_{\sigma(1)} - \eta - 2s_0^\kappa > 0$,

$$R_{\tilde{\pi}_\eta}(T) \leq \sum_{k=2}^{K}[s_0 + 2\int_0^\infty \exp(-au^{1+\kappa p/p-1})du](\mu_{\sigma(1)} - \theta_i)$$

The range constraint of $p \in (1, 2]$ in Theorem 2 can be easily addressed: if we know $\mathbb{E}[|X_k|^p] \leq u$ for a certain $p > 2$, then $\mathbb{E}[|X_k|^2] \leq u + 1$.

Proofs

Proof of Theorem 20:

Let event $\mathcal{E}(t)$ be true if an arm is selected at time t in its round robin turn. For any $k \neq \sigma(1)$ and $t > K$:

$$\{\pi_\eta(t) = k\} \subseteq \{\pi_\eta(t) = k, \overline{\mathcal{E}}(t)\} \cup \{\pi_\eta(t) = k, \mathcal{E}(t)\}.$$

Let $\bar{\mu}_{k,s}$ denote the sample mean of arm k calculated from its first s observations. For the total number of times that the first event can happen we have,

$$\mathbb{E}\sum_{t=K+1}^{\infty}\mathbb{I}\{\pi_\eta(t) = k, \overline{\mathcal{E}}(t)\} \leq \mathbb{E}\sum_{t=K+1}^{\infty}\mathbb{I}\{\pi_\eta(t) = k, \bar{\mu}_k(t) > \eta + \tau_k(t)^{-\frac{1}{3}}\}$$

$$
\begin{aligned}
&\leq \mathbb{E}\sum_{s=1}^{\infty}\mathbb{I}\{\bar{\mu}_{k,s} > \eta + s^{-\frac{1}{3}}\} \\
&\leq \sum_{s=1}^{\infty}\mathbb{P}\{\bar{\mu}_{k,s} > \eta + s^{-\frac{1}{3}}\} \\
&\leq \sum_{s=1}^{\infty}\mathbb{P}\{\bar{\mu}_{k,s} - \mu_k > s^{-\frac{1}{3}}\} \qquad (7.6)\\
&\leq \sum_{s=1}^{\infty}\exp(-as^{\frac{1}{3}}) \qquad (7.7)\\
&\leq \int_0^{\infty}\exp(-au^{\frac{1}{3}})du \\
&= \frac{6}{a^3},
\end{aligned}
$$

where 7.6 holds because $\mu_i \leq \eta$ and 7.7 holds by 3.3. For the second event in 7.6 we have

$$
\begin{aligned}
\mathbb{E}\sum_{t=K+1}^{\infty}\mathbb{I}\{\pi_\eta(t) = k, \mathcal{E}(t)\} &\leq \mathbb{E}\sum_{s=1}^{\infty}\mathbb{I}\{\bar{\mu}_{\sigma(1),s} < \eta + s^{-\frac{1}{3}}\} \qquad (7.8)\\
&\leq \sum_{s=1}^{\infty}\mathbb{P}\{\bar{\mu}_{\sigma(1),s} < \eta + s^{-\frac{1}{3}}\} \\
&\leq s_0 - 1 + \sum_{s=s_0}^{\infty}\mathbb{P}\{\bar{\mu}_{\sigma(1),s} < \eta + s^{-\frac{1}{3}}\} \\
&\leq s_0 - 1 + \sum_{s=s_0}^{\infty}\mathbb{P}\{\bar{\mu}_{\sigma(1),s} - \mu_{\sigma(1)} < -(\mu_{\sigma(1)} - \eta - s_0^{-\frac{1}{3}})\} \\
&\leq s_0 - 1 + \sum_{s=s_0}^{\infty}\exp(-a(\mu_{\sigma(1)} - \eta - s_0^{-\frac{1}{3}})^2 s) \qquad (7.9)\\
&\leq s_0 - 1 + \int_0^{\infty}\exp(-a\delta_0^2 u)du \\
&\leq s_0 - 1 + \frac{1}{a\delta_0^2}
\end{aligned}
$$

for any $s_0 \in \mathbb{N}$ such that

$$
\delta_0 \triangleq \mu_{\sigma(1)} - \eta - s_0^{-1/3} > 0. \qquad (7.10)
$$

The inequality 7.8 holds because, for $t > K$ for any round robin cycle to start, the inequality $\bar{\mu}_{\sigma(1),s} < \eta + s^{-\frac{1}{3}}$ needs to hold for some s. The inequality 7.9 holds

by 3.3. Thus

$$R_{\pi_\eta}(T) \leq \frac{6}{a^3}(\mu_{\sigma(1)} - \eta)(K-1) + \sum_{k=2}^{K}[s_0 + \frac{1}{a\delta_0^2}](\mu_{\sigma(1)} - \mu_{\sigma(k)}).$$

for any s_0 and δ_0 satisfying 7.10. ■

Proof of Theorem 21:

Similar to the proof of the Theorem 20, define $\mathcal{E}(t)$. Following the similar lines as in the proof of Theorem 20, we have

$$\begin{aligned}
\mathbb{E}\sum_{t=K+1}^{\infty}\mathbb{I}\{\widetilde{\pi}_\eta(t) = k, \mathcal{E}(t)\} &\leq \mathbb{E}\sum_{t=K+1}^{\infty}\mathbb{I}\{\pi_\eta(t) = k, \tilde{\mu}_k(t) > \eta + \tau_k(t)^\kappa\} \\
&\leq \mathbb{E}\sum_{t=K+1}^{\infty}\mathbb{I}\{\tilde{\mu}_{k,s} > \eta + s^\kappa\} \\
&\leq \mathbb{E}\sum_{s=1}^{\infty}\mathbb{I}\{\tilde{\mu}_{k,s} > \eta + 4u^{1/p}(\frac{\log(\epsilon_s^{-1})}{s})^{\frac{p-1}{p}}\} \\
&\leq \sum_{s=1}^{\infty}\mathbb{P}\{\tilde{\mu}_{k,s} - \mu_{\sigma k} > 4u^{1/p}(\frac{\log(\epsilon_s^{-1})}{s})^{\frac{p-1}{p}}\} \\
&\leq \sum_{s=1}^{\infty}\exp(-ak^{1+\kappa p/p-1}) \qquad (7.11) \\
&\leq \int_0^{\infty}\exp(-au^{1+\kappa p/p-1})du \qquad (7.12)
\end{aligned}$$

where 7.11 is by 10.5. Also,

$$\begin{aligned}
\mathbb{E}\sum_{t=+1}^{\infty}\mathbb{I}\{\pi_\eta(t) = i, \overline{\mathcal{E}}(t)\} &\leq \mathbb{E}\sum_{s=1}^{\infty}\mathbb{I}\{\tilde{\mu}_{\sigma(1),s} < \eta + s^\kappa\} \\
&\leq \mathbb{E}\sum_{s=1}^{\infty}\mathbb{I}\{\tilde{\mu}_{\sigma(1),s} - \mu_{\sigma(1)} < -(\mu_1 - \eta - s^\kappa)\} \\
&\leq s_0 - 1 + \mathbb{E}\sum_{s=s_0}^{\infty}\mathbb{I}\{\tilde{\mu}_{\sigma(1),s} - \mu_{\sigma(1)} < -s^\kappa\} \\
&\leq s_0 - 1 + \sum_{s=s_0}^{\infty}\mathbb{P}\{\tilde{\mu}_{\sigma(1),s} - \mu_{\sigma(1)} < -s^\kappa\}
\end{aligned}$$

$$
\begin{aligned}
&\leq s_0 - 1 + \sum_{s=s_0}^{\infty} \mathbb{P}\{\tilde{\mu}_{\sigma(1),s} - \mu_{\sigma(1)} < -4u^{1/p}(\frac{\log(\epsilon_s^{-1})}{k})^{\frac{p-1}{p}}\} \\
&\leq s_0 - 1 + \sum_{s=s_0}^{\infty} \exp(-ak^{1+\kappa p/p-1}) \\
&\leq s_0 - 1 + \int_0^{\infty} \exp(-au^{1+\kappa p/p-1})du. \qquad (7.13)
\end{aligned}
$$

We arrive at the theorem by 7.12 and 7.13. ■

7.2 Conclusion

For the classic MAB problem it has been shown that the regret grows at least in a logarithmic order with time. The unbounded growth of regret with time shows that the mistakes in arm selection occur infinitely often and any uniformly good policy never converges to the best arm. However, given some side information, bounded regret, thus complete learning, can be achieved. In this chapter, we developed online learning policies with bounded regret that use less amount of side information comparing with the existing work.

Part II

Sequential Inference

CHAPTER 8

LITERATURE REVIEW ON SEQUENTIAL INFERENCE

In this section we review the classic results on sequential inference. Provided sequences of observations from an environment the objective is to detect particular underlying phenomenons with the smallest possible number of observations. The essence of the problems is the tension between the delay and the reliability: the desired reliability can be achieved through the accumulation of measurements, which comes at the price of increasing the detection delay.

The sequential inference problems considered here can be categorized into three classes of problems: sequential hypothesis testing, change detection, and active hypothesis testing.

8.1 Sequential Hypothesis Testing

The classic sequential hypothesis testing problem was pioneered by Wald [6]. The observation sequence $\{X(t)\}_{t=1,\ldots,\infty}$ is generated identically ad independently according to a distribution f_0 or f_1 depending on whether H_0 or H_1 is the true hypothesis (underlying state of nature). The goal is to design a sequential policy that at each time decides *i)* whether to continue observations or not *ii)* wether to declare H_0 or H_1 as the true hypothesis. Specifically, a sequential test $\pi = (\tau, \delta)$ consists of a stopping time τ and a terminal decision δ. After observation of τ samples, one of the two hypotheses is declared as the true one. Let $\delta = 0$ denote the declaration of hypothesis H_0 and $\delta = 1$ denote the declaration of hypothesis H_1 . Particularly, the objective is to minimize the expected sample number, $\mathbb{E}[\tau]$,

subject to the following constraints on the probability of error

$$\mathbb{P}[\delta = 1|H_0] \leq \alpha, \tag{8.1}$$

$$\mathbb{P}[\delta = 0|H_1] \leq \beta, \tag{8.2}$$

for small positive α and β. The first type of error given in 8.1 is referred to as false alarm and the second type of error given in 8.2 is referred to as missed detection.

Wald showed that the sequential probability ratio test (SPRT) is asymptotically optimal for this problem. In SPRT the sampling is stopped at time

$$\tau = \min_{t \geq 1}\{R_t \geq A \text{ or } R_t \leq B\} \tag{8.3}$$

where $R_t = \Pi_{s=1}^{t} \frac{f_1(X(s))}{f_0(S(s))}$ is the likelihood ratio and $A > 1 > B > 0$ are the stopping boundaries. The decision is $\delta = \mathbb{I}\{R_\tau \geq A\}$. The thresholds A and B are designed such that the error probability constraints are met: $\alpha = \mathbb{P}_{f_0}[R_\tau \geq A]$ and $\beta = \mathbb{P}_{f_1}[R_\tau \leq B]$. Calculating the exact values of A and B is quite laborious. Instead of exact values of A and B, the so called Wald's approximation values can be used in practice. The Wald's approximations of the values are

$$A = \frac{1-\beta}{\alpha}, \quad B = \frac{\beta}{1-\alpha}.$$

Under the above choice of A and B Wald showed that

$$\mathbb{E}_{f_0}[\tau] = \frac{-\alpha \log A - (1-\alpha)\log B}{I(f_0, f_1)},$$
$$\mathbb{E}_{f_1}[\tau] = \frac{(1-\beta)\log A + \beta \log B}{I(f_1, f_0)},$$

under H_0 and H_1, respectively. The sample complexity of SPRT is asymptotically optimal as α and β go to 0.

In the simple hypothesis testing, it is assumed that the distributions are fully known under two hypotheses. The variations of this problem where the distributions are unknown up to a parameter θ have been studied in the literature referred to as composite hypothesis testing problems. The composite hypotheses testing problem is fundamentally more difficult than simple hypothesis testing. The sequential generalized likelihood ratio test (SGLRT) was first studied by Schwartz for one-parameter exponential family with i.i.d. observations [66]. Specifically, consider the problem of testing $H_0 : \theta < \theta_0$ versus $H_1 : \theta > \theta_1$ for the exponential family of distributions with parameter θ ($\theta_1 > \theta_0$). Schwartz proposed to test the maximum likelihood estimate $\hat{\theta}_t$ of the unknown parameter at each time t against the boundary parameters θ_0 and θ_1. Under a Bayesian formulation of the objective function with cost 1 for rejecting the true hypothesis and cost c per observation, Schwartz test is asymptotically optimal as $c \to 0$. A refinement of [66] was studied by Lai [67, 68] which showed that for a multivariate exponential family, SGLRT asymptotically optimizes the Bayesian cost even with no indifference region (when $\theta_0 = \theta_1$).

Another well-studied test for sequential composite hypothesis testing is the sequential adaptive likelihood ratio test (SALRT) [69, 70, 71]. While SALRT has computationally more efficient statistics, its poor early estimates can never be revised even with a large number of observations. All these classic results assume i.i.d. observations over time.

The optimality of SPRT for sequential hypothesis testing with non-stationary observations was shown in [72]. The optimal SPRT in the non-stationary environment requires laborious calculation of a sequence of thresholds. The asymptotic optimality of SPRT with approximated thresholds, under certain assump-

tions on log-likelihood ratios, was shown in [73, 74].

8.2 Quickest Change-Point Detection

A closely related problem to hypothesis testing is the change-point detection. This problem was first studied by Shewart in 1931 [75] for the application of online quality control of a manufacturing process. The conventional setting of quickest change-point detection involves a random process $\{X(t)\}_{t\geq 1}$ in which the observations $\{X(t)\}_{t=1}^{\nu-1}$ before an unknown change point ν are i.i.d with distribution f_0 and the observations $\{X(t)\}_{t\geq \nu}$ after the change point are i.i.d with distribution f_1. The objective is to detect the change point ν as quickly as possible subject to a reliability constraint. The essence of the problem is the tension between the objective and the constraint: the desired reliability can be achieved through the accumulation of measurements, which comes at the price of increasing the detection delay.

There are two standard formulations of the change-point detection problem: Bayesian and frequentist. The Bayesian formulation was pioneered by Shiryaev in 1960's [7, 8], where the change point ν is assumed to be a random variable with a known distribution and the objective is to minimize the expected detection delay subject to a constraint on the probability of false alarm. Under the frequentist (i.e., minimax) formulation, the unknown change point ν is deterministic. This work focuses on the Bayesian formulation.

The objective in the Bayesian quickest change-point detection is to design a

stopping time τ that minimizes the expected detection delay

$$\mathbb{E}[(\tau - \nu)^+] \tag{8.4}$$

under a constraint on the probability of false alarm

$$\mathbb{P}[\tau < \nu] \leq \alpha. \tag{8.5}$$

Based on a recursive dynamic programming approach and assuming a geometric distribution for the change point with i.i.d. pre- and post-change observations, Shiryaev showed in [7] that the optimal solution to this problem is a threshold policy. The Shiryaev test τ_S is the first time instant that the posterior probability that change has happened exceeds a certain threshold:

$$\tau_S = \min\{n \in \mathbb{N} : \mathbb{P}[\nu \leq n | X(1), X(2), ..., X(n)] \geq A\}. \tag{8.6}$$

The threshold A is chosen such that $\mathbb{P}[\tau_S < \nu] = \alpha$. While the calculation of the exact threshold is laborious, a simple conservative choice of $A = 1 - \alpha$ satisfies the false alarm constraint and offers asymptotic optimality as α approaches 0.

A generalization of Shiryaev's results to non-i.i.d. observations was given in [76]. It was shown in [76] that the Shiryaev test given in 8.6 is asymptotically optimal under this general model provided that the average log-likelihood ratio (LLR) converges. Specifically, it is assumed in [76] that the time-varying pre-change and post-change distributions $\{f_{0,t}\}_{t=1}^{\nu-1}$ and $\{f_{1,t}\}_{t\geq\nu}$ are such that the average LLR

$$\frac{1}{n-k}\sum_{s=k}^{n} \log \frac{f_{1,s}(Y(s))}{f_{0,s}(Y(s))} \tag{8.7}$$

converges to a positive constant q almost surely as n goes to infinity for any starting time instant k. This assumption preserves the linear relationship between the sum LLR and the detection delay as in the original i.i.d. case. This

linear relationship was exploited in [76] to obtain probabilistic bounds on the detection delay given bounds on the sum LLR (and vice versa) in establishing the asymptotic optimality of the Shiryaev test.

8.3 Active Inference

The problem of sequential design of experiments where a set of different experiments are available and the observations depend on the chosen experiment was studied by Chernoff [9]. Different from the the sequential hypothesi testing problems described above, the Chernoff's sequential design of experiments allows the decision maker to interactively choose the experiments. Under each experiment, observations are generated from different distributions for each hypothesis. This model referred to as active hypothesis testing (also controlled sensing for hypothesis testing) thus has another dimension of interactively choosing the more informative experiments. Chernoff introduced a randomized test that optimizes the Bayesian cost for active hypothesis testing. Cohen and Zhao [77] introduced a simpler deterministic test that offer asymptotic optimality while having a better performance in finite regime.

CHAPTER 9

SEQUENTIAL HYPOTHESIS TESTING AND CHANGE DETECTION UNDER TIME-VARYING MODELS

In this chapter, we consider sequential hypothesis testing problems under exponentially time-varying distribution models. In particular, the decision maker aims to infer the state of an underlying phenomenon from a sequence of observations. The observation sequence available to the decision maker $\{X(t)\}_{t=1}^{\infty}$ takes the general form of a random process with exponentially time varying mean: $X(t) = \exp(\theta t) + N(t)$. The state of the underlying phenomena is encoded in the value of θ unknown to the decision maker. The decision maker only has access to the noisy observations of $\exp(\theta t)$ with an additive i.i.d. noise $N(t)$. The objective is to minimize the inference delay subject to an error probability constraint.

We analyze the sequential hypothesis testing problem under two settings referred to as binary hypothesis testing and quickest change point detection.

In the simplest formulation of the above problem, referred to as binary hypothesis testing, the observations are drawn from two different random processes determined by two different values of θ depending on whether hypothesis H_0 or H_1 is true. The objective is to minimize the detection delay subject to error probability constraints.

In the second setting, referred to as quickest change point detection, the observations before an unknown change point ν are drawn from a random process determined by a parameter θ_0 and after the change point are drawn from a random process determined by a parameter θ_1.The objective is to detect the change

point ν as quickly as possible subject to a reliability constraint. In other words, the decision maker aims at choosing one of the hypotheses $\{H_i\}_{i=1}^{\infty}$ with the smallest number of observations where hypothesis H_i indicates $\nu = i$.

The essence of the problem is the tension between the objective and the constraint: the desired reliability can be achieved through the accumulation of measurements, which comes at the price of increasing the detection delay.

In this chapter, we develop asymptotically optimal tests for the above problems under different scenarios where the problem parameters are known (simple testing) or unknown (composite testing). Specifically, we develop detection tests and analyze their performance. Moreover, by providing analysis for the fundamental performance limitations we prove the asymptotic optimality of the proposed tests.

The results find applications in instability detection of a general linear system with distinct real eigenvalues as well as nonlinear systems as shown later in the motivating example.

Notation

Let us introduce some concepts and notations that are used throughout the chapter. Let $X^{(t)} = X(1), X(2), ..., X(t)$, and $f(X^{(t)}; \theta)$ denote the joint distribution of $X^{(t)}$. The notation $l_t(\theta_1, \theta_0)$ denotes the log-likelihood ratio of two distributions with parameters θ_1 and θ_0 at time t,

$$l_t(\theta_1, \theta_0) = \log \frac{f(X(t); \theta_1 t)}{f(X(t); \theta_0 t)}. \tag{9.1}$$

The Kullback-Leibler (KL) divergence between the above two distributions, denoted by $I_t(\theta_1, \theta_0)$ is defined as the following expectation of the log-likelihood

ratio.

$$I_t(\theta_1, \theta_0) = \mathbb{E}_{\theta_1} l_t(\theta_1, \theta_0), \tag{9.2}$$

where $\mathbb{E}_\theta[.]$ is the expectation operator when θ is the parameter determining the underlying distributions.

9.1 Motivation: Voltage Instability Detection in Power Systems

Voltage stability in a power system refers to the ability of the system to maintain the load voltage within specified operating limits. The voltage stability problem is classified into short-term and long-term stability phenomena [78]. Short-term voltage instability phenomenon is mainly caused by heavy usage of reactive power by electronically controlled loads and induction motors.

Short-term instability can be characterized by the system Lyapunov exponents [79]. In particular, short-term voltage instability occurs if one of the Lyapunov exponents is positive. To detect voltage instability, it is therefore natural to use the Lyapunov exponents or a proxy of Lyapunov exponents as the indicator of instability. Existing techniques estimate Lyapunov exponents (or related statistics) from phasor measurement unit (PMU) data or state estimates [79, 80]. These existing techniques are heuristic and do not provide any level of performance guarantee.

Lyapunov Exponents

The Lyapunov exponents in a non-linear system are analogous to the eigenvalues of a linear system in the sense that they carry information about the stability of the system. The Lyapunov exponents of a non-linear system are defined as follows [81].

Definition 1 *Consider a continuous-time dynamical system* $\dot{x} = f(x)$ *with* $x \in \mathcal{X} \in \mathbb{R}^n$. *Let* $\psi(t, x)$ *be the solution of the differential equation. Define the following limiting matrix*

$$\Gamma(x) = \lim_{t\to\infty}[\frac{\partial\psi(t,x)^T}{\partial x}\frac{\partial\psi(t,x)}{\partial x}]^{\frac{1}{2t}}. \tag{9.3}$$

Let $\Lambda_i(x)$ *be the eigenvalues of the limiting matrix* $\Gamma(x)$. *The Lyapunov exponents* $\lambda_i(x)$ *are defined as*

$$\lambda_i(x) = \log \Lambda_i(x). \tag{9.4}$$

Without loss of generality, assume that $\lambda_1(x) \geq \lambda_2(x) \geq ... \geq \lambda_n(x)$, where $\lambda_1(x)$ is referred to as the maximum Lyapunov exponent.

An algorithm for online computation of Lyapunov exponents with improved computational efficiency was proposed in [80]. Let $V_{m\Delta} \in \mathbb{R}^n$ be the measured voltage at time $m\Delta$, where $m = 0, 1, ..., M$ and $\Delta > 0$ is the sampling interval. Choose $0 < \epsilon_1 < \epsilon_2$ and an integer L such that $\epsilon_1 < \|V_{m\Delta} - V_{(m-1)\Delta}\| < \epsilon_2$ for $m = 1, 2, ..., L$. The maximum Lyapunov exponent λ_1 at time $k\Delta$ is obtained as follows [80]. For $k > L$,

$$\exp(Lk\Delta\lambda_1) = \Pi_{m=1}^{L}\frac{\|V_{(k+m)\Delta} - V_{(k+m-1)\Delta}\|}{\|V_{m\Delta} - V_{(m-1)\Delta}\|}. \tag{9.5}$$

Let $\theta = L\Delta\lambda_1$ and $t = k - L$. Let $X(t)$ denote the statistic obtained from measurements (the right hand side of equation 9.5), referred to as the sample observation. Taking into account the effect of noise, we can write the sample observations of a power system at time t in the form of

$$X(t) = e^{\theta t} + N(t), \tag{9.6}$$

where the noise $N(t)$ is assumed to be normally distributed. By definition, θ indicates the stability of the system: the system is stable when $\theta < 0$ and instable when $\theta > 0$.

9.2 Binary Hypothesis Testing: Problem Formulation

The problem considered in this work can be formulated as a sequential hypothesis testing problem with time-varying distribution of observations. In particular, under each hypothesis, observations are ruled by a non-stationary random process determined by a parameter θ. The null hypothesis corresponds to the stable system, $H_0 : \theta < 0$. The alternative hypothesis corresponds to the instable system, $H_1 : \theta > 0$. The objective, similar to the classic sequential hypothesis testing problem, is to minimize the expected sample number subject to the error constraints. To start with, we assume that, under each hypothesis, the parameter is known (See Sec. 9.3.1). In practical applications, however, the value of parameters may be unknown. In Sec. 9.3.2 , we study the sequential detection problem under exponentially time-varying distribution model with unknown parameters.

In our formulation, the distribution of the sample observation at time t is

$$f(X(t);\theta t) = \frac{1}{\sqrt{2\pi}} \exp(\frac{(X(t) - e^{\theta t})^2}{2}), \tag{9.7}$$

where θ is the true parameter. Note that the distribution of sample observations is time-varying within a specific model of exponential dependence to the parameter.

9.3 Binary Hypothesis Tests

In this section, we propose asymptotically optimal tests for the time-varying hypothesis testing under both simple and composite hypotheses. We show the asymptotic optimality by analyzing the performance of the tests and providing performance limitation for any test under this setting.

9.3.1 Under Simple Hypothesis Model

To gain insight into the similarities and differences between our problem and the classic binary hypothesis testing, we first consider the simple hypothesis case where $H_0 : \theta = \theta_0 < 0$ and $H_1 : \theta = \theta_1 > 0$. The constraint on the first and second type of error is given by α and β, respectively, as in 8.1 and 8.2.

Define

$$L_t(\theta_1, \theta_0) = \sum_{s=1}^{t} l_s(\theta_1, \theta_0), \tag{9.8}$$

$$g_{\theta_1,\theta_0}(t) = \sum_{s=1}^{t} I_s(\theta_1, \theta_0). \tag{9.9}$$

and the inverse function $g^{-1}_{\theta_1,\theta_0}(z)$

$$g^{-1}_{\theta_1,\theta_0}(z) \triangleq \min\{t \in N : g_{\theta_1,\theta_0}(t) \geq z\}. \tag{9.10}$$

The SPRT-exp, a modification of the classic SPRT, is as follows. Continue sampling as long as

$$\log B < L_t(\theta_1, \theta_0) < \log A, \tag{9.11}$$

stop sampling otherwise. The terminal decision is given by

$$\delta^{SPRT-exp} = \begin{cases} 0, & if \;\; L_t(\theta_1, \theta_0) \leq \log B, \\ 1, & if \;\; L_t(\theta_1, \theta_0) \geq \log A. \end{cases} \tag{9.12}$$

The thresholds A and B are designed such that the error probability constraints are met. Calculating the exact values of A and B is extremely laborious. Instead of exact values of A and B, the so called Wald's approximation values can be used in practice. The Wald's approximations of the values are

$$A = \frac{1-\beta}{\alpha}, \quad B = \frac{\beta}{1-\alpha}.$$

An upper bound on the expected number of observations for the SPRT-exp is given in the following theorem.

Theorem 22 *The expected number of observations for the* $\pi^{SPRT-exp} = (\tau^{SPRT-exp}, \delta^{SPRT-exp})$ *satisfies*

$$\mathbb{E}_{\theta_0}[\tau^{SPRT-exp}] \leq g^{-1}_{\theta_1,\theta_0}(-(1-\alpha)\log B - \alpha \log A), \tag{9.13}$$

$$\mathbb{E}_{\theta_1}[\tau^{SPRT-exp}] \leq g^{-1}_{\theta_1,\theta_0}((1-\beta)\log A + \beta \log B). \tag{9.14}$$

proof 17 *See Appendix B.*

Recall that in the classic simple hypothesis testing considered by Wald it was shown that the average sample number equals to

$$\begin{aligned}\mathbb{E}_{f_0}[\tau] &= \frac{-(1-\alpha)\log B - \alpha \log A}{I(f_0, f_1)},\\ \mathbb{E}_{f_1}[\tau] &= \frac{(1-\beta)\log A + \beta \log B}{I(f_1, f_0)},\end{aligned}$$

A comparison between Theorem 22 and the classic problem shows a logarithmic order of reduction in the average sample number for our test.

9.3.2 Under Composite Hypothesis Model

Dictated by the practical constraints, the exponent parameters are often unknown to the decision maker. We thus introduce and analyze the performance of SGLRT-exp for the composite hypothesis testing problem. Conventionally, the set of possible parameters Θ is partitioned to three disjoint sets. Under hypothesis H_0, $\theta \in \Theta_0$, under hypothesis H_1, $\theta \in \Theta_1$, where $\Theta_0 \cap \Theta_1 = \emptyset$, and $\mathcal{I} = \Theta/\{\Theta_0 \cup \Theta_1\} \neq \emptyset$ is an indifference set. In this problem the indifference set is assumed to be the $(-a, a)$ interval for some small $a > 0$. The sets Θ_0 and Θ_1 are the $(-d, -a)$ and (a, d) intervals, respectively. The Bayes cost assigns cost one for the declaration of hypothesis H_1 (or H_0) when hypothesis H_0 (or H_1) is the true one. Also, obtaining each sample observation incurs a cost of $c > 0$. The objective of a sequential test is to minimize the Bayes cost which is equivalent to

$$R_0^\pi = c\mathbb{E}_\theta[\tau] + \mathbb{P}_\theta[\delta = 1] \quad or \tag{9.15}$$

$$R_1^\pi = c\mathbb{E}_\theta[\tau] + \mathbb{P}_\theta[\delta = 0], \tag{9.16}$$

when the true parameter θ is in Θ_0 or Θ_1, respectively. The $\pi^{SGLRT-exp}$ is as

follows. At time t calculate a maximum likelihood estimation of the parameter,

$$\widehat{\theta_t} = \arg \sup_{\theta \in \Theta/\mathcal{I}} f(X^{(t)}; \theta). \tag{9.17}$$

For any $\theta \in \Theta/\mathcal{I}$ define $\rho(\theta)$ the index of the alternative hypothesis of θ. In other words if $\theta \in \Theta_0$ let $\rho(\theta) = 1$, otherwise, if $\theta \in \Theta_1$ let $\rho(\theta) = 0$. Calculate the ϕ_t as

$$\phi_t = \arg \sup_{\theta \in \Theta_{\rho(\widehat{\theta_t})}} f(X^{(t)}; \theta). \tag{9.18}$$

Continue observation of new samples as long as

$$L_t(\widehat{\theta_t}, \phi_t) < -\log c, \tag{9.19}$$

stop observation, otherwise. The terminal decision is given by

$$\delta^{SGLRT-exp} = 1 - \rho(\widehat{\theta_\tau}). \tag{9.20}$$

Next, we establish an upper bound on the performance of SGLRT-exp. Moreover, we provide a lower bound on the performance of any arbitrary sequential composite hypothesis test of exponents that shows the asymptotic optimality of SGLRT-exp. It is assumed the true parameter is $\theta_0 \in \Theta_0$. The similar results hold if the alternative hypothesis is the true one. For any $\theta \in \Theta/\mathcal{I}$, let

$$\psi(\theta) = \arg \inf_{\psi \in \Theta_{\rho(\theta)}} I_1(\theta, \psi). \tag{9.21}$$

Also, let $g_{\theta_0}(t) = g_{\theta_0, \psi(\theta_0)}(t)$, accordingly, $g^{-1}_{\theta_0}(z) = g^{-1}_{\theta_0, \psi(\theta_0)}(z)$.

Theorem 23 *The Bayes cost of the SGLRT-exp satisfies*

$$R_0^{SGLRT-exp} \leq (1 + \epsilon) c \, g^{-1}_{\theta_0}(-\log c), \tag{9.22}$$

such that $\epsilon \to 0$ *as* $c \to 0$.

proof 18 *See Appendix B*

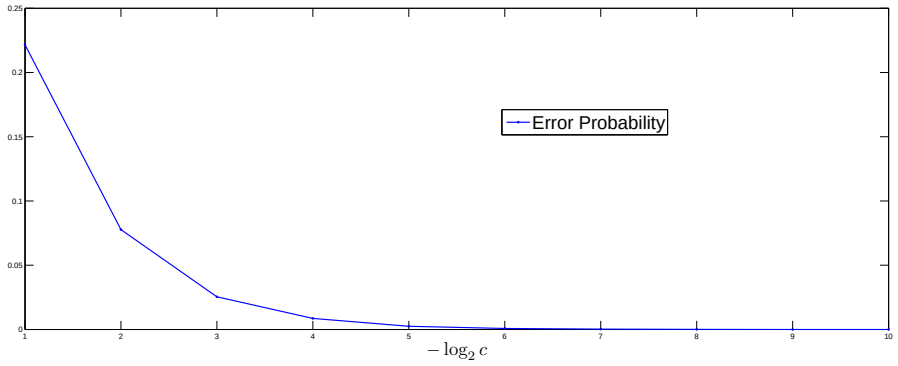

Figure 9.1: The probability of error for SGLRT-exp.

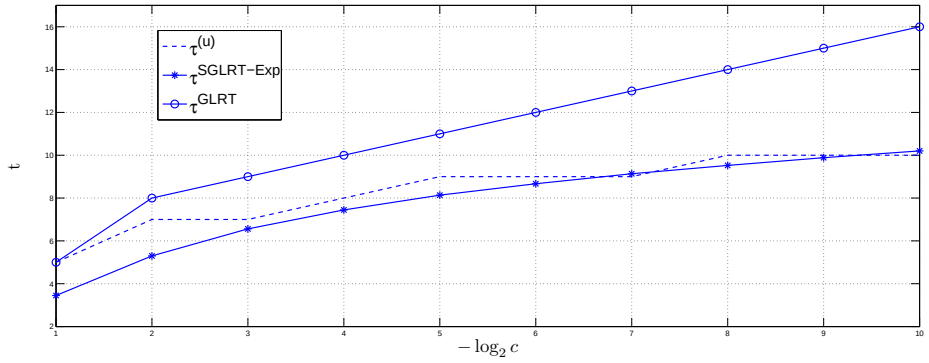

Figure 9.2: The average sample number.

Theorem 24 *The Bayes cost of any sequential hypothesis test of exponents π satisfies*

$$R_0^{\pi} \geq (1-\epsilon)c\, g_{\theta_0}^{-1}(-(1-\epsilon)\log c), \tag{9.23}$$

such that $\epsilon \to 0$ as $c \to 0$.

proof 19 *See Appendix B.*

9.3.3 Simulations

In this section, numerical analysis of the performance of the SGLRT-exp is provided. The numerical results as we shall see are close to asymptotic upper bounds provided in the work. Furthermore, to show the efficiency of the sequential test, the average sample number of the SGLRT-exp is compared with a fixed size test with the same power. Fig. 9.1 shows the probability of error for the SGLRT-exp over different values of the cost c. For smaller c, the cost of obtaining observations is lower, thus a higher number of observations results in a smaller probability of error. Second figure, shows the average sample number for the SGLRT-exp algorithm ($\tau^{SGLRT-exp}$). From our analytical results in Theorem 23, the value of $\tau^{(u)} = g_{\theta_0}^{-1}(-\log c)$ is an approximation of the upper bound on the average sample number, which is illustrated in the figure. Also, denoted by τ^{GLRT}, the number of observations required in a fixed size GLRT to achieve the same probability of error is shown in Fig. 9.2 that confirms the efficiency of the sequential algorithm. In these simulations the values of d and a are assigned to 1 and 0.05, respectively. For the first two figures $\theta_0 = 0.1$. The second two figures show the same numerical analysis when $\theta_0 = -0.1$.

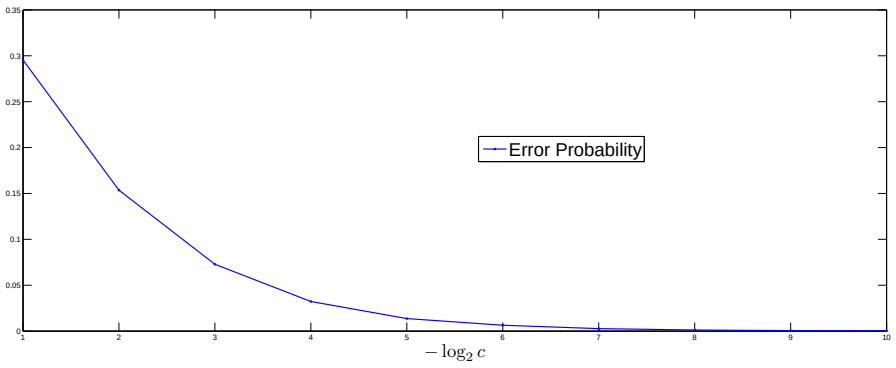

Figure 9.3: The probability of error for SGLRT-exp.

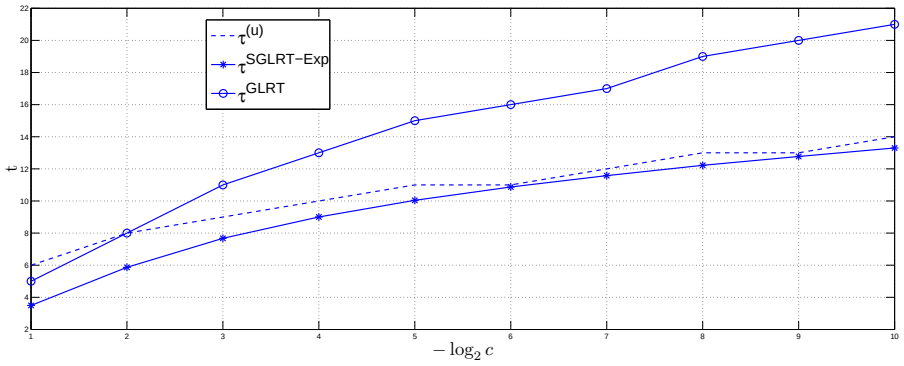

Figure 9.4: The average sample number.

9.4 Bayesian Quickest Change-point Detection: Problem Formulation

The quickest change detection problem studied in this work can be formulated as a quickest change-point detection problem where the pre-change and post-change observations are ruled by non-stationary random processes. In our formulation, the distribution of the sample observation at time t is

$$f(X(t);\theta_t t) = \frac{1}{\sqrt{2\pi}} \exp(\frac{(X(t) - e^{\theta_t t})^2}{2}), \tag{9.24}$$

where $\theta_t = \theta_0 < 0$ for $t < \nu$ (the system is stable) and $\theta_t = \theta_1 > 0$ for $t \geq \nu$ (the system becomes instable). Note that the distribution of the observations is time-varying with a specific model of exponential dependence to the parameters. We point out that the assumption on the average LLR (see 8.7) adopted in [76] does not hold in this problem where the sum LLR grows exponentially. In particular, the relation between the expected sum LLR and the detection delay is strictly convex rather than linear. As a consequence, different techniques are needed in establishing asymptotically optimal tests.

We assume that the change-point distribution $p_k = \mathbb{P}[\nu = k]$ satisfies the exponential tail condition

$$\exp(-\xi_1 n) \leq \mathbb{P}[\nu > n] \leq \exp(-\xi_2 n). \tag{9.25}$$

The objective is the same as in the classic change-point detection given in 8.4 and 8.5.

Before the random instability change point ν, the system operates under nominal conditions. The system parameter θ_0 can be estimated. We thus assume that the pre-change parameter θ_0 is known. For practical applications, however, the value of post-change parameter θ_1 may be unknown. We study both cases of known and unknown θ_1.

Let $\mathbb{P}^{(k)}$ and $\mathbb{E}^{(k)}$ denote, respectively, the probability and expectation operators for $\nu = k$.

9.5 Quickest Change-Point Detection Tests

In this section, we propose asymptotically optimal tests for the time-varying change-point detection under both known and unknown post-change parameter. We show the asymptotic optimality by analyzing the performance of the tests and providing performance limitation for any test under this setting.

9.5.1 Under Known Post-Change Parameter

We first consider the case when the post-change parameter θ_1 is known. We show that the Shiryaev test is asymptotically optimal.

As introduced in [76], define

$$\omega_n(\theta_1, \theta_0) = \frac{\mathbb{P}[\nu \leq n | X^{(n)}]}{\mathbb{P}[\nu > n | X^{(n)}]}, \tag{9.26}$$

where the argument (θ_1, θ_0) may be omitted when it is clear from the context. Then Shiryaev test with a threshold of A can be rewritten as

$$\tau_S = \min\{n \in \mathbb{N} : \omega_n \geq \gamma\}, \tag{9.27}$$

where $\gamma = \frac{A}{1-A}$. Let

$$L_k^n(\theta_1, \theta_0) = \sum_{s=k}^{n} l_s(\theta_1, \theta_0), \tag{9.28}$$

and

$$g_{\theta_1,\theta_0}(n; k) = \sum_{s=k}^{n} I_s(\theta_1, \theta_0) \tag{9.29}$$

denote, respectively, the sum LLR and the sum KL divergence from time k to n. Also define the inverse function $g^{-1}_{\theta_1,\theta_0}(z; k)$ for a fixed starting time k as

$$g^{-1}_{\theta_1,\theta_0}(z; k) \triangleq \min\{n \in N : g_{\theta_1,\theta_0}(n; k) \geq z\}. \tag{9.30}$$

For the change-point detection problem, the relation between the statistic ω_n and the sum LLR can be easily obtained as

$$\omega_n = \frac{1}{\mathbb{P}[\nu > n]} \sum_{k=1}^{n} p_k e^{L_k^n}. \tag{9.31}$$

We show next that the Shiryaev test with a fixed threshold $\gamma = \frac{1-\alpha}{\alpha}$ is asymptotically optimal as α approaches 0. This result is obtained by establishing an upper bound on the expected delay of the Shiryaev test (Theorem 25) that matches with a lower bound on the expected delay of any test (Theorem 26) as α approaches 0.

Theorem 25 *For the change-point detection problem, there exists $\epsilon > 0$ such that the expected delay of the Shiryaev test with $\gamma = \frac{1-\alpha}{\alpha}$ satisfies*

$$\mathbb{E}[(\tau_S - \nu)|\tau_S > \nu] \leq (1+\epsilon) \sum_{k=1}^{\infty} p_k g^{-1}_{\theta_1,\theta_0}(\log(1/\alpha); k), \tag{9.32}$$

where $\epsilon \to 0$ as $\alpha \to 0$.

proof 20 *Based on the value of the threshold γ and the definition of the Shiryaev test τ_S, we have*

$$\omega_{\tau_S - 1} \leq \frac{1-\alpha}{\alpha}. \tag{9.33}$$

The inequalities 9.31 and 9.33 lead to an upper bound on $L_k^{\tau_S - 1}$. The next step is to derive an upper bound on $\mathbb{E}^{(k)}[\tau_S - k]$ from the upper bound on $L_k^{\tau_S - 1}$. In [76], the proof for this step relies on the assumption on the convergence of the average LLR given in 8.7. Without this assumption, our proof hinges on establishing an upper bound on $\mathbb{P}^{(k)}[\tau_S - k - g^{-1}_{\theta_1,\theta_0}(\log \frac{1-\alpha}{\alpha}, k) \geq i]$ for all positive i. The upper bound on $\mathbb{E}[\tau_S - \nu]$ given in 9.32 is then obtained from the upper bounds on $\mathbb{E}^{(k)}[\tau_S - k]$. The detailed proof is given in Appendix B.

The following theorem gives a lower bound on the expected delay of any test τ.

Theorem 26 *For any test for the instability detection problem satisfying $\mathbb{P}[\tau < \nu] \leq \alpha$, there exists $\epsilon > 0$ such that*

$$\mathbb{E}[\tau - \nu|\tau \geq \nu] \geq (1-\epsilon) \sum_{k=1}^{\infty} p_k g^{-1}_{\theta_1,\theta_0}(\log(1/\alpha); k) \tag{9.34}$$

where $\epsilon \to 0$ as $\alpha \to 0$.

proof 21 *Let* $\eta_{\alpha,\epsilon}(k) = g^{-1}_{\theta_1,\theta_0}((1-\epsilon)\log(1/\alpha); k)$ *for some* $\epsilon > 0$. *By Markov inequality*

$$\mathbb{E}^{(k)}[(\tau - k)^+] \geq \eta_{\alpha,\epsilon}(k)\mathbb{P}^{(k)}[\tau - k \geq \eta_{\alpha,\epsilon}(k)]. \tag{9.35}$$

The desired lower bound on $\mathbb{E}[\tau - \nu | \tau > \nu]$ *is then obtained from 9.35 after showing* $\mathbb{P}^{(k)}[k \leq \tau < k + \eta_{\alpha,\epsilon}(k)] \to 0$ *as* $\alpha \to 0$. *The detailed proof is given in Appendix B.*

Theorems 25 and 26 together prove the asymptotic optimality of Shiryaev test with a fixed threshold $\gamma = \frac{1-\alpha}{\alpha}$ for the instability detection problem.

9.5.2 Under Unknown Post-Change Parameter

We now consider the quickest change point detection problem with an unknown post-change exponent. We develop an asymptotically optimal test which can be viewed as the Shiryaev test with a maximum likelihood (ML) estimate of the post-change parameter θ_1.

Assume that $\theta_0 \leq -\delta$ for some small positive δ and $\theta_1 > \delta$ belongs to a discrete set Θ_1 with a finite cardinality $|\Theta_1|$. The objective is a test that minimizes the expected detection delay under the constraint that the false alarm probability is capped below α for all $\theta_1 \in \Theta_1$.

The proposed test under an unknown post-change parameter uses the following statistic:

$$\widetilde{\omega}_n = \sup_{\theta \in \Theta_1} \omega_n(\theta, \theta_0). \tag{9.36}$$

Specifically, the statistic $\widetilde{\omega}_n$ is obtained by substituting the ML estimate of θ_1 into $\omega_n(\theta_1, \theta_0)$. Note that an upper bound on $\widetilde{\omega}_n$ is also an upper bound on $\omega_n(\theta_1, \theta_0)$

for all $\theta_1 \in \Theta_1$. The proposed test is then given by a stopping time defined by comparing $\tilde{\omega}_n$ with a time-varying threshold γ_n:

$$\tau_{ML} = \min\{n \in \mathbb{N} : \widetilde{\omega}_n \geq \gamma_n\}. \tag{9.37}$$

In order to satisfy the condition on probability of false alarm, threshold γ_n in τ_{ML} need to be set at a higher value comparing to the Shiryaev test. In Shiryaev test, $\omega_n(\theta_1, \theta_0) \geq \frac{1-\alpha}{\alpha}$ translates directly to the condition on probability of false alarm. Here, since $\widetilde{\omega}_n \geq \omega_n(\theta_1, \theta_0)$, the value of threshold needs to be designed more carefully. Theorem 27 shows that with $\gamma_n = \frac{|\Theta_1|\mathbb{E}[\nu]}{\alpha\mathbb{P}[\nu>n]}$ the proposed test offers asymptotic optimality. Note that the asymptotic performance proven in this theorem holds uniformly for all values of $\theta_1 \in \Theta_1$.

Theorem 27 *For the instability detection problem with an unknown post-change parameter, the false alarm probability and the expected detection delay of the proposed test τ_{ML} with $\gamma_n = \frac{|\Theta_1|\mathbb{E}[\nu]}{\alpha\mathbb{P}[\nu>n]}$ satisfy the following:*

$$\begin{aligned} \mathbb{P}[\tau_{ML} < \nu] &\leq \alpha, \\ \mathbb{E}[\tau_{ML} - \nu | \tau_{ML} \geq \nu] &\leq (1+\epsilon) \sum_{k=1}^{\infty} p_k g^{-1}_{\theta_1,\theta_0}(\log(1/\alpha); k). \end{aligned}$$

where $\epsilon \to 0$ as $\alpha \to 0$.

proof 22 *The upper bound on the false alarm probability is obtained by a change of probability measure using the sum LLR statistic. Noticing that*

$$\omega_{\tau_{ML}-1}(\theta_1, \theta_0) \leq \widetilde{\omega}_{\tau_{ML}-1} \leq \gamma_{\tau_{ML}-1}, \tag{9.38}$$

we can establish the upper bound on the expected detection delay with a similar line of arguments as in the proof of Theorem 25. The detailed proof is given in Appendix B.

Since knowing the value of the post-change parameter will not increase the best possible detection delay, the same lower bound as given in Theorem 26 holds

for the case of unknown post-change parameter. Theorem 3 thus establishes the asymptotic optimality of the proposed test τ_{ML}.

9.5.3 Simulations

For the simulation results given here, we assume that the change point ν follows a geometric distribution with $p_k = \rho(1-\rho)^{k-1}$. The pre- and post-change parameters are given as $\theta_0 = -0.1$ and $\theta_1 = 0.1$. The constraint on the false alarm probability is set to $\alpha = 2^{-i}$ with $i = 4, 5, ..., 14$, to illustrate the performance of the proposed tests.

In Figures 9.7 and 9.8, we plot the expected detection delay of τ_S and τ_{ML} (assuming θ_1 is unknown and $\theta_1 \in \Theta_1 = \{0.01, 0.02, 0.03..., 0.5\}$) as a function of $-\log\alpha$. Figures 9.7 and 9.8 show a higher expected delay for τ_{ML} comparing to τ_S. Figures 9.5 and 9.6 show the false alarm probabilities of τ_S and τ_{ML} for different geometric distributions of the change point. We see from Figures 9.5 and 9.6 that the both tests are conservative in terms of satisfying the false alarm constraint. It is the same with the Shiryaev test when the threshold is set at $1-\alpha$. The τ_{ML} obtains a lower false alarm probability comparing to τ_S. Higher expected delay and lower false alarm probability in τ_{ML} is because of the higher thresholds in τ_{ML}.

9.6 Instability Detection in General Linear Systems

The results developed in the previous sections apply to the quickest detection of instability in a first-order linear system in the presence of noise. The solution

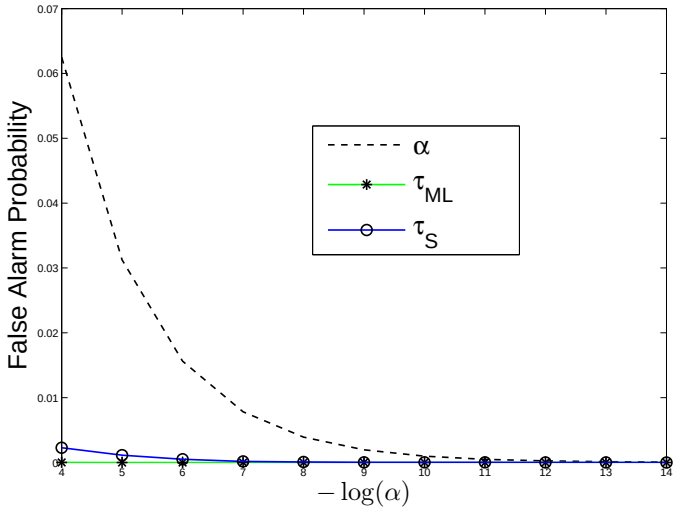

Figure 9.5: Probability of error. $\rho = 0.01$.

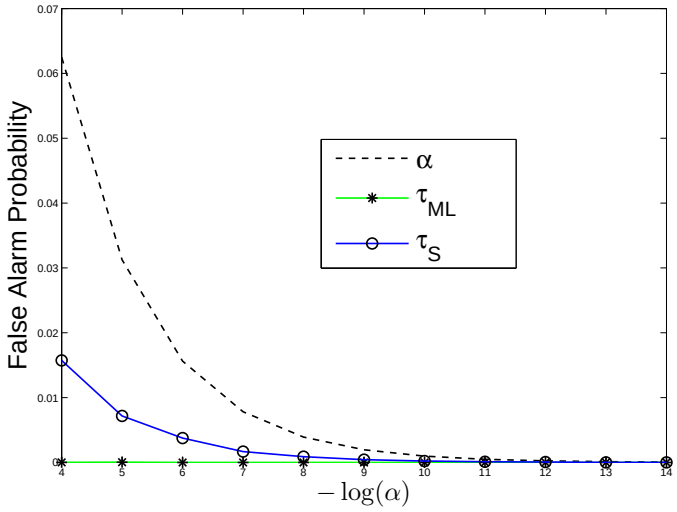

Figure 9.6: Probability of error. $\rho = 0.1$.

to a K-dimensional first-order discrete-time linear system under the assumption of distinct real eigenvalues is given by

$$y(t) = \sum_{i=1}^{K} a_i e^{\lambda_i t} + N(t), \tag{9.39}$$

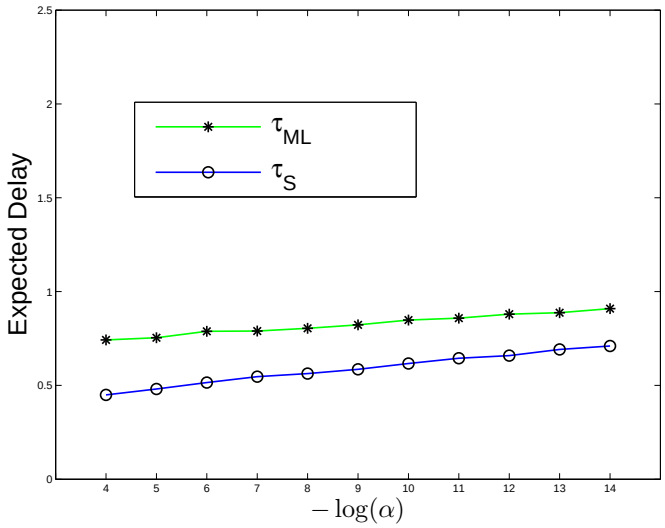

Figure 9.7: Expected voltage instability detection delay. $\rho = 0.01$.

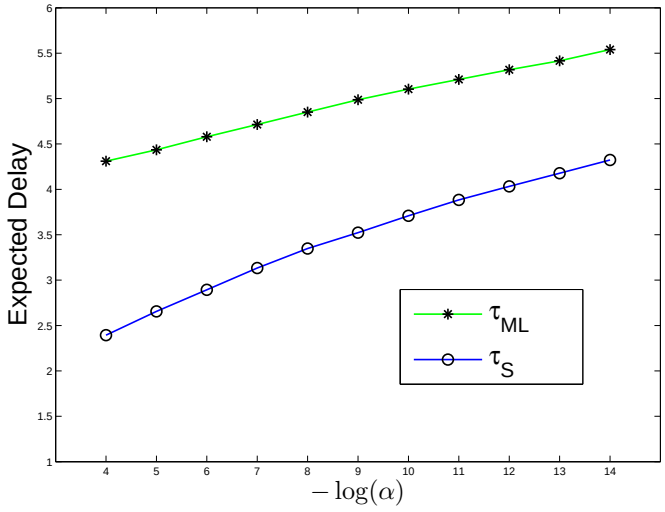

Figure 9.8: Expected voltage instability detection delay. $\rho = 0.1$.

where λ_i denotes the eigenvalues, $N(t)$ the noise, and $a_i \in \mathbb{R}$ the coefficients. The system is instable if and only if, at least, one of the eigenvalues is positive. The quickest detection of the instability of a linear system can thus be similarly formulated as the voltage instability detection problem. Let Λ =

$(a_1, a_2, ...a_k; \lambda_1, \lambda_2, ...\lambda_k)$ denote the set of parameters representing the solution to the system. Let $f(y(t); \Lambda, t)$ be the distribution of $y(t)$. Similar to Section 5.3, we can define sum LLR and the sum of KL-divergence as

$$L_k^n(\Lambda_1, \Lambda_0) = \sum_{s=k}^{n} \log \frac{f(y(s); \Lambda_1, s)}{f(y(s); \Lambda_0, s)}, \tag{9.40}$$

$$g_{\Lambda_1,\Lambda_0}(n; k) = \sum_{s=k}^{n} \mathbb{E}_{\Lambda_1} \log \frac{f(y(s); \Lambda_1, s)}{f(y(s); \Lambda_0, s)}. \tag{9.41}$$

Since the exponential term corresponding to the largest eigenvalue is dominant the results apply with straightforward modifications.

When the system has repetitive or complex eigenvalues, the quickest instability detection problem is more involved and requires further study.

9.7 Conclusion

In this chapter, we proposed asymptotically optimal tests for hypothesis testing problems within a particular time variation model. In our model, the mean value of the random process is allowed to change exponentially over time. This formulation is applicable to instability detection in linear systems as well as non-linear systems as demonstrated in the example of Lyapunov exponents in power system.

CHAPTER 10

ACTIVE INFERENCE UNDER HIERARCHICAL OBSERVATIONS AND UNKNOWN MODELS

We consider the problem of detecting a few targets among a large number of hierarchical data streams. The data streams are modeled as a set of stochastic processes with unknown and potentially heavy-tailed distributions. The stochastic nature of the data streams may result from the inherent randomness of the underlying phenomenon or the noisy response of the measuring process.

Target data streams manifest themselves in their abnormal mean values. More specifically, a data stream is a target of interest if its mean value exceeds a certain prescribed threshold. There is partial knowledge on the ordering of the data streams in terms of their mean values, and such prior knowledge is assumed to induce a tree-structured hierarchy. As illustrated in Figure 10.1 for the special case of a binary-tree hierarchy, each node in the tree represents a data stream, and the parent-children relation encodes the known ordering in that the mean value of the parent is no smaller than the maximum mean value of its children. The objective is to detect all targets quickly and reliably by fully exploiting the tree-encoded prior knowledge. Specifically, we seek an active inference strategy that determines, sequentially, which node on the tree to probe in order to minimize the sample complexity for achieving a given level of

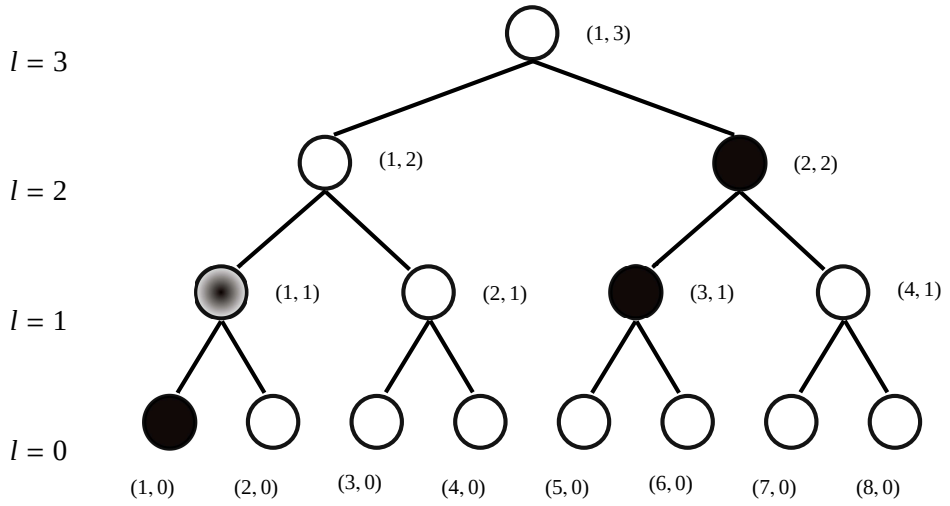

Figure 10.1: The hierarchical data streams model.

detection reliability.

10.0.1 Applications

The above general problem of detecting abnormal mean values with partial ordering knowledge arises in a number of active inference and learning applications in networking and data analytics. We give below several representative examples.

Heavy hitter and hierarchical heavy hitter detection: In Internet and other communication and financial networks, it is common that a small number of flows, referred to as heavy hitters, account for the most of total traffic [82]. Heavy hitters are naturally defined as flows with an abnormal mean value above a certain threshold. The hierarchical heavy hitter (HHH) detection problem has also been studied as a variation that extends the notion of heavy hitter to hierarchical data structures. Specifically, HHH is defined as a heavy hitter whose mean value remains above the threshold after excluding its heavy hit-

ter decedents. This problem finds applications in pinpointing denial-of-service (DoS) attacks and Distributed DoS attacks [83], discovering worms [84] and other anomalies.

The recent improvements in software defined networking (SDN) allows programmable routers to count aggregated flows that match a given IP prefix [85] which induces a binary-tree structured hierarchy. The partial ordering in packet counts at different IP prefixes naturally follows from that the aggregated flow has a high packet count per unit time than each of its constituent flows.

Noisy group testing: In group testing, the objective is to identify a small subset of defective items within a large set of items. The problem pioneered by Dorfman [86], allows for the test of multiple items where the output of a test is positive or negative depending on the presence or absence of the defective items. This problem finds a variety of applications in industrial and medical diagnosis, and anomaly detection.

The classic group testing focuses on noiseless test outcomes and ignores the measurement noise that can be considerable specially when larger sets of the items are tested. See related work in Sec. 10.0.3. In a noisy group testing problem, the test outcomes can be flipped with a possibly unknown error probability.

In the noisy group testing, where the targets are the defective items and the data streams are the noisy responses to the group tests, a nested test plan induces a tree structured hierarchy. The mean value of the test responses for each group including a target is above a known threshold. The partial ordering is a result of the fact that if a set contains a target its supersets contain at least a target as well.

Adaptive sampling with sequential queries: Consider the problem of adaptive sampling for estimation of a step function in $[0, 1]$ interval provided sequential queries about the location of the step with noisy responses. The problem finds applications in active learning of binary threshold classifiers which arise for example in document classification [87], as well as stochastic root finding [88].

Partitioning the $[0, 1]$ interval to smaller intervals, the hierarchical target search problem applies to the adaptive design of the inquiries, where the target is a small interval containing the location of the step. Inquiries about larger intervals (consisting of a number of smaller intervals) induces a hierarchical structure for the responses. Partial ordering is a result of the fact that an interval whose a sub-interval contains the target, also contains the target.

10.0.2 Main Results

We develop an active inference strategy which adaptively chooses the sampling points at each time and determines the targets at the time of stopping. The proposed strategy, referred to as Confidence Bound based Random Walk (CBRW), induces a random walk on the tree. Each step in the random walk is taken according to the comparison between the confidence bounds on the mean value of the processes (obtained from sequential observations) and the thresholds determining the anomalous behavior. The progress is designed in a way that at each step the random walk is more likely to get closer to a target rather than away from it.

Appropriate for different applications, we consider two different settings of

leaf-level targets and hierarchical targets. In the first setting, the targets are specified to be the processes with a mean value above a predetermined threshold at the leaf level. In CBRW, once the current location of the random walk is on a leaf node the corresponding data stream will be examined for being a target. In the second setting, the targets are any node in the tree whose mean value is above a threshold after excluding anomalous decedents. We introduce a variation of the CBRW which examines all the nodes in the tree for being a hierarchical target in an iterative fashion.

The proposed search strategy is computationally efficient since they only comprise of calculating the confidence bounds for the mean values of data streams and performing simple comparisons.

We analyze the performance of the strategies and show a logarithmic-order sample complexity in the number K of data streams provided that the gap in the mean value of the observations and the threshold at all levels of the tree is bounded away from 0 and can be treated as a constant. Such gaps show the informativeness of the observations at different levels of the tree. The logarithmic-order performance in K is optimal as determined by information theoretic lower bound. We further analyze the reliability of the target set found by the search strategies. We show that the sample complexity grows in a logarithmic-order with $\frac{1}{\epsilon}$ when the confidence level is set at $1-\epsilon$. The $O(\log K + \log \frac{1}{\epsilon})$ sample complexity for the CBRW is a result of the random walk structure of the strategy which efficiently separates the objectives of getting closer to the target in $O(\log K)$ observations and determining the target confidently in $O(\log \frac{1}{\epsilon})$ observations.

The results on the target search problem are obtained under two general set-

tings for the unknown distribution model: *i)* Sub-Gaussian distributions and *ii)* Heavy-Tailed distributions with bounded b'th moment ($1 < b \leq 2$). The second setting covers a very large class of distributions and the only requirement is the existence of the b'th moment. For example, this setting covers all distributions with bounded variance. Different settings for the distribution model are suitable for different applications.

10.0.3 Related Work

The target search problem in hierarchical data streams studied in this work is related to several active learning and sequential inference problems as discussed.

Variations of the classic group testing have been extensively studied in the literature. Most of the existing work focuses on the noisless test outcomes and the non-adaptive methods which translates the problem into a classic coding problem [89, 90]. The problem of characterizing the number of tests in terms of reliability constraint is not considered in the literature. However, there are several recent works that consider one-sided error in the tests (false positive or false negative) [91, 92, 93] or symmetric error (with equal false positive and false negative probabilities) [94, 95, 96]. In these existing works the error probabilities (in test outcomes) is assumed to be known a priori and is used in the policy design. While, our approach to the problem allows an arbitrary unknown error probability.

The main body of the results on the adaptive sampling with noisy observations is based on a Bayesian approach with a binary noise. In the Bayesian approach, a priori distribution is assumed for z^*, and the posterior distribution

of z^* is updated after each observation. The Bayesian approach assumes fixed and known distribution of the noise which is required in the Bayesian update of the posterior distribution of z^*. In the popular Bayesian strategy, referred to as Probabilistic Bisection Algorithm (PBA), the sampling point at each time is chosen to be the median point of the distribution of z^*. Although several variations of the method have been extensively studied in the literature [97, 98, 99] after the pioneering work of [100], there is little known about the theoretical guarantees according to [99], specially when it comes to unknown distribution of the noise. In this paper we present a non-Bayesian approach to the adaptive sampling problem under general models for noise with unknown distribution.

Prior solutions for online detection of HHHes typically involve adjusting which prefixes to monitor either (1) based on the arrival of each packet [101, 102, 103], or (2) at a periodic interval [104, 105]. The former would require custom hardware to cope with high link speed. In the latter approach, prefixes are adjusted based on measurements of a periodic interval and can be efficiently implemented in commodity hardware such as ternary content addressable memory (TCAMs). Our model is similar to the adaptive monitoring algorithm presented by Jose et al [104], where a fixed number of measurement rules N are adjusted at periodic intervals based on the aggregate count of packets matching to each of these rules. At each time interval, the absolute aggregate count is compared to a threshold (e.g., fraction of link capacity), to determine if it is an HHH, and whether the rules need to be kept in the next interval, or expanded to monitor the children of the prefix, or collapsed and combined with upstream nodes. Although our model is similar, the decision criteria used to adjust the prefix is different. Instead of comparing with a fixed threshold, our decision is based on a statistical metric bounded by the desired detection er-

ror. We provide a more rigorous framework that succinctly capture the tradeoff between detection time and overall detection performance.

A similar problem of target search with hierarchical observations was first studied in [106]. A sequential random walk policy was developed in [106] based on sequential probability ratio tests (SPRT) assuming a known distribution model. Under known distribution models the problem can be formulated as an active sequential simple hypothesis testing. In this work, different from [106], we study the target search problem under unknown distribution models. Our work can be seen as an active sequential non-parametric composite hypothesis testing which uses inherently different techniques in designing the steps of the random walk as well as in analysis.

The target search problem studied here is related to the Pure Exploration Bandits [107] where the objective is to search for a subset of bandit arms with certain properties. In particular, in the Thresholding Bandit problem introduced in [108], the objective is to determine the arms with mean above a given threshold. The thresholding bandit similarly finds a various number of applications in industrial production, crowdsourcing, active and discrete level set detection, and active classification. The focus of our work is however on exploiting the hierarchical structure to obtain sublinaer sample complexity with the problem size.

10.1 Problem Formulation

Consider a set of data streams modeled as a set of i.i.d. random processes with unknown distributions. There is partial knowledge on the ordering of the data

streams in terms of their mean values, and such prior knowledge is assumed to induce a tree-structured hierarchy. Each node in the tree represents a data stream, and the parent-children relation encodes the known ordering in that the mean value of the parent is no smaller than the maximum mean value of its children.

For the simplicity of presentation of the main techniques and ideas in this work, we start with a binary-tree structured data as shown in Fig. 10.1, with $K = 2^L$ ($L \in \mathbb{N}$) leaf nodes. We show the extension to more general data structures in Sec. 10.4. We use a pair (k, l) to index the nodes in the tree where l denotes the level as shown in Fig. 10.1. Let $\{X_{k,l}(t)\}_{t=1}^{\infty}$ denote the corresponding random processes and $\mathcal{F} = \{f_{k,l}\},\ l = 0, ..., L,\ k = 1, ...2^{L-l}.$ denote the unknown distribution model. We use the terms node, random process and data stream interchangeably. Assume a threshold $\eta_l \in \mathbb{R}$ corresponds to each level l of the tree.

There is a set of targets defined as the random processes with anomalous mean behavior to be detected. We consider two different settings:

In the first setting, the targets are the leaf nodes whose mean value is above a given threshold $\eta_0 \in \mathbb{R}$. Let $\mathcal{S} = \{k : \mu_{k,0} > \eta_0\}$ denote the set of such targets, where $\mu_{k,l}$ is the mean value of process (k, l). We further assume that the mean value of the parents of the target nodes at level l are above η_l. Specifically, $\mu_{k,l} > \eta_l$ if and only if the node (k, l) is on the shortest path from a target node to the root node.

In the second setting, the hierarchical targets are defined as each node (k, l) on $\mathcal{T}$ (not necessarily the leaf nodes) whose mean value remains above the

threshold η_l after excluding its anomalous decedents. The anomalous decedents are all decedents of (k, l) whose mean value are above the threshold at their respective level.

In the example shown in Fig. 10.1, node $(1, 0)$ is a target in the first setting. Also, nodes $(1, 0)$, $(3, 1)$, $(2, 2)$ are examples of hierarchical targets. There is a possibility that a node with anomalous mean (e.g. node $(1, 1)$) has a mean value below the threshold (η_1) after excluding its anomalous decedent (node $(1, 0)$). In that case, this node is not considered as a hierarchical target, although it has an anomalous mean.

The goal is to design a sequential strategy $\pi = (\{a_t\}_{t\geq 1}, T_\pi, \delta)$ to interactively select the sampling points a_t at each time $1 \leq t \leq T_\pi$, and declare the set of targets $\mathcal{S}_\delta$ at the time of stopping T_π. The objective is to minimize the expected number of samples $\mathbb{E}_\mathcal{F}[T_\pi]$ subject to a reliability constraint $\mathbb{P}_\mathcal{F}[\mathcal{S}_\delta \neq \mathcal{S}] \leq \epsilon$, for $\epsilon \in (0, 0.5)$. Notations $\mathbb{E}_\mathcal{F}$ and $\mathbb{P}_\mathcal{F}$, respectively, denote the expectation and probability under distribution model $\mathcal{F}$. Formally,

$$\begin{aligned} &\text{minimize}_\pi \ \mathbb{E}_\mathcal{F} T_\pi, \\ &s.t. \ \ \mathbb{P}_\mathcal{F}[\mathcal{S}_\sigma \neq \mathcal{S}_\eta] \leq \epsilon. \end{aligned}$$

The policy design is without the knowledge of the distributions except for some very general assumptions. We consider the problem under two distribution models. We present the results under Sub-Gaussian distribution model. We then discuss the extension to heavy-tailed distributions in Sec. 10.4. Recall that a real-valued random variable X is called Sub-Gaussian if it satisfies the following [29],

$$\mathbb{E}[e^{u(X-\mathbb{E}[X])}] \leq e^{\xi u^2/2} \tag{10.1}$$

for some constant $\xi > 0$. For Sub-Gaussian random variables, Hoeffding-like concentration inequalities hold. Specifically [30]:

$$\begin{aligned} \mathbb{P}\Big[\overline{X}(s) + \sqrt{\frac{2\xi \log \frac{1}{p}}{s}} < \mu\Big] &\leq p, \\ \mathbb{P}\Big[\overline{X}(s) - \sqrt{\frac{2\xi \log \frac{1}{p}}{s}} > \mu\Big] &\leq p. \end{aligned} \tag{10.2}$$

For heavy-tailed distributions, upper bounds on moment generating function (similar to 10.1) no longer exist. However, with the assumption of an upper bound on the moments of order $1 < b \leq 2$, we still can obtain similar confidence intervals. In particular, if for a random variable X

$$\mathbb{E}[X^b] \leq u, \tag{10.3}$$

for some $u > 0$, we can use the truncated sample mean defined as

$$\widehat{X}(s,p) = \frac{1}{s}\sum_{t=1}^{s} X(t)\mathbb{1}\left\{|X(t)| \leq (\frac{ut}{\log \frac{1}{p}})^{1/b}\right\} \tag{10.4}$$

to obtain confidence intervals on mean value of the random proceses. Particularly for any $p \in (0, \frac{1}{2}]$,

$$\begin{aligned} \Pr\left[\widehat{X}(s,p) - 4u^{1/b}(\tfrac{\log \frac{1}{p}}{s})^{\frac{b-1}{b}} > \mu\right] &\leq p \\ \Pr\left[\widehat{X}(s,p) + 4u^{1/b}(\tfrac{\log \frac{1}{p}}{s})^{\frac{b-1}{b}} < \mu\right] &\leq p. \end{aligned} \tag{10.5}$$

For a proof for 10.5, see Lemma 1 in [31].

10.2 An Active Inference Strategy: CBRW

In this section, we propose a sequential target search strategy referred to as Confidence Bounds based Random Walk (CBRW). We first focus on the case of a

single target and sub-Gaussian distributions. Extensions to multiple target detection and heavy-tailed distribution models are discussed in Sec. 10.4.

10.2.1 Detecting Leaf-Level Targets

The basic structure of CBRW consists of a global random-walk module interwoven with a local CB-based sequential test at each step of the random walk. Specifically, the CBRW policy performs a biased random walk on the tree that eventually arrives and terminates at the target with the required reliability. Each move in the random walk (i.e., which neighboring node to visit next) is guided by the output of the local CB-based sequential tests. The local CB-based sequential sampling test ensures that the random walk is more likely to move toward the target than to move away from the target and that the random walk terminates at the true target with high probability.

Let us first specify the local CB-based sequential sampling test $\mathcal{A}$. The goal of the test is to determine wether the mean value of a random process is above or below a predetermined threshold at certain confidence levels. Test $\mathcal{A}$ sequentially collects samples from the process and calculates upper and lower confidence bounds for the mean value of the process at each time. The sampling stops when the upper confidence bound goes below the threshold or the lower confidence bound goes above the threshold. Specifically,

$$\overline{X}(s) + \sqrt{\frac{2\xi \log \frac{2s^3}{p_1}}{s}} < \eta, \text{ or} \tag{10.6}$$

$$\overline{X}(s) - \sqrt{\frac{2\xi \log \frac{2s^3}{p_2}}{s}} > \eta, \tag{10.7}$$

where $1 - p_1$ $(1 - p_2)$ is the confidence level for the upper (lower) confidence bound, $\overline{X}(s)$ is the sample mean obtained form s observations, and ξ is the distribution parameter specified in 10.1. We assume (an upper bound on) ξ is known. If 10.6 is satisfied, the mean value of the random process is below η with probability $1 - p_1$; and if 10.7 is satisfied, the mean value of the random process is above η with probability $1 - p_2$. In the former case, we assign 0 to the output of test $\mathcal{A}$: $o_{\mathcal{A}_{p_1,p_2,\eta}}(X) = 0$ and in the latter case we assign 1 to the output of test $\mathcal{A}$: $o_{\mathcal{A}_{p_1,p_2,\eta}}(X) = 1$, indicating likeliness of absence and presence of target, respectively.

The global random walk on the tree is guided by the outputs of the local CB-based tests. In particular, let a pointer (k, l) denote the current location of the random walk (which is initially set at the root node). Left child of (k, l) is tested according to $\mathcal{A}$. If the output is 1 (indicating the likeliness of presence of target in the branch whose root node is the left child), the pointer is moved to the left child. Otherwise, the right child is tested similarly and if the output is 1, the pointer is moved to the right child. If the output of test $\mathcal{A}$ on both children is 0 the pointer is moved to the parent of the current location (the parent of the root node is defined as itself). The first time that the output of test $\mathcal{A}$ is 1 at a leaf node, CBRW stops and declares the leaf node as the target.

The parameters of the local CB-based test are designed as follows. For the test of non-leaf nodes set: $p_1 = p_2 = p_0$ where $p_0 \in (0, 1 - \frac{1}{\sqrt{2}})$, and the threshold $\eta = \eta_l$ at the respective level. The choice of parameters ensures getting closer to the target with a probability more than 0.5. For the test of leaf nodes set: $p_1 = p_0$ and $p_2 = \frac{\epsilon}{2LC_{p_0}}$ and $\eta = \eta_0$ where the value of C_{p_0} (a constant independent of K and ϵ) is specified in 10.13. The choice of parameters ensures detecting the

target at the desired confidence level as proven in Theorem 28.

10.2.2 Detecting Hierarchical Targets

A variation of CBRW detects the hierarchical targets at the desired confidence level. The difference from CBRW presented in previous subsection is in the design of the steps in the random walk and the criteria in declaring the target. Specifically, in this variation, the steps of the random walk are designed as follows. The current location of the random walk is tested according to $\mathcal{A}$ with parameters $p_1 = p_2 = p$, where $p = p_0 \in (0, 1 - \frac{1}{\sqrt[3]{2}})$, and the threshold $\eta = \eta_l$ at the respective level. If the output is 0, the pointer is moved to the parent of the current location. If the output of test $\mathcal{A}$ at the current location of the pointer is 1, then the children are tested one by one and if one of them is likely to contain the target according to $\mathcal{A}$, the pointer is moved to that child. In the case that the output of test $\mathcal{A}$ at the current location of the pointer is 1, and the output of the test at both children of the current node is 0, the current location is likely to be a hierarchical target. The current location of the pointer is thus further examined for being a hierarchical target by diving p by 2 (increasing the confidence level) and repeating the local CB-based tests at the current location of the pointer and its children. If the the same results are obtained for the output of the tests at the current location of the pointer and its children, the process with be repeated with p again divided by 2. If the value of p goes below $\frac{\epsilon}{3LC^H_{p_0}}$, the current location of the pointer is declared as a hierarchical target. The value of $C^H_{p_0}$ is specified in 10.15 and ensures detecting the target at the desired confidence level as proven in Theorem 29.

10.3 Performance Analysis

In this section, we provide analysis for the sample complexity of the proposed sequential target search policy. We establish an optimal logarithmic-order upper bound in both K and $\frac{1}{\epsilon}$ under both leaf-level and hierarchical target settings. The analysis focuses on the case of a single target and Sub-Gaussian distributions. Extensions to multiple target detection and heavy-tailed distribution models are discussed in Sec. 10.4.

We start the analysis of CBRW by providing an upper bound on the sample complexity of the local CB-based test $\mathcal{A}$ in Lemma 9. The sample complexity of test $\mathcal{A}$ is different on target and non-target nodes since the parameters of the test are tuned differently. We then establish an upper bound on the sample complexity of CBRW in Theorems 28 and 29 under both settings of leaf-level and hierarchical targets. The proof of Theorems 28 and 29 is based on analyzing the trajectory of the pointer to have an upper bound on the number of times which random processes are tested according to $\mathcal{A}$.

Lemma 9 *Let μ be the expected value of an i.i.d. Sub-Gaussian random process $\{X(t)\}_{t=1}^{\infty}$ and $\overline{X}(s)$ be the sample mean obtained from the first s sequential observations. Stopping time T is defined as the first time in which one of the following events happens*

$$T = \min\Big\{s : \overline{X}(s) + \sqrt{\frac{2\xi \log \frac{2s^3}{p_1}}{s}} < \eta, \text{ or } \overline{X}(s) - \sqrt{\frac{2\xi \log \frac{2s^3}{p_2}}{s}} > \eta\Big\}. \tag{10.8}$$

We have, if $\mu > \eta$,

$$\mathbb{P}[\overline{X}(T) + \sqrt{\frac{2\xi \log \frac{2T^3}{p_1}}{T}} < \eta] \leq p_1, \text{ and} \tag{10.9}$$

$$\mathbb{E}[T] \leq \frac{48}{(\mu-\eta)^2} \log \frac{24\sqrt[3]{\frac{2}{p_2}}}{(\mu-\eta)^2} + 2. \tag{10.10}$$

Also if $\mu < \eta$,

$$\mathbb{P}[\overline{X}(T) - \sqrt{\frac{2\xi \log \frac{2T^3}{p_2}}{2T}} > \eta] \leq p_2, \text{ and} \tag{10.11}$$

$$\mathbb{E}[T] \leq \frac{48}{(\mu-\eta)^2} \log \frac{24\sqrt[3]{\frac{2}{p_1}}}{(\mu-\eta)^2} + 2. \tag{10.12}$$

proof 23 *See Appendix for the proof.*

10.3.1 Leaf-Level Target Setting

We first introduce some auxiliary notions which are useful in understanding the trajectory of the random walk in CBRW. Consider a sequence of subtrees $\{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_L\}$ of $\mathcal{T}$. Subtree $\mathcal{T}_L$ is obtained by removing the biggest half-tree containing the target from $\mathcal{T}$. Subtree $\mathcal{T}_l$ is iteratively obtained by removing the biggest half-tree containing the target from the half-tree containing the target in the previous iteration. The subtrees $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ are shown in Fig. 10.2 for a $\mathcal{T}$ with $K = 8$ where the node $(1, 0)$ is assumed to be the target. Let (k_l, l) denote the child of the root node of $\mathcal{T}_l$.

The gap $\mu_{k,l} - \eta_l$ in the mean value of a random process at level l and the threshold at the respective level indicates the informativeness of the observations. We naturally assume that the higher levels are less informative; thus, have smaller gaps. For example, in group testing, tests from larger groups of items are less informative about the presence of defective items. The constants

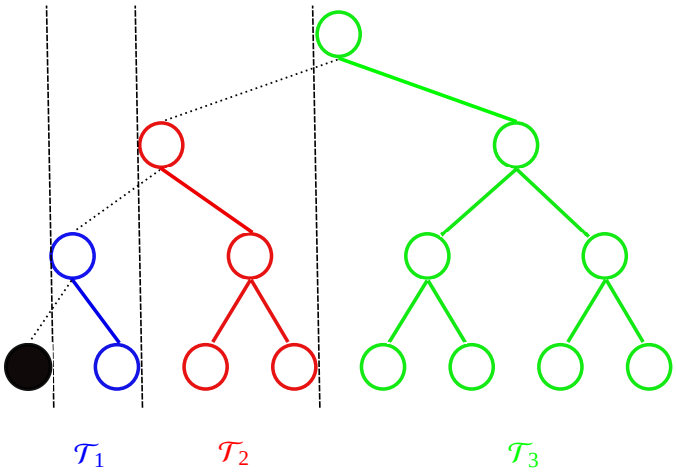

Figure 10.2: The hierarchical data streams model.

in Theorems 28 and 29 are obtained under this assumption. While, the logarithmic sample complexity in K and $\frac{1}{\epsilon}$ holds regardless of this assumption.

Theorem 28 *Provided that the gap in the mean value of the processes and the threshold is bounded away from zero at all levels of the tree, the sample complexity of CBRW is in optimal logarithmic order in both K and $\frac{1}{\epsilon}$:*

$$\mathbb{E}_{\mathcal{F}}[T_{RW-CB}] = O(\log_2 K + \log \frac{1}{\epsilon}).$$

A detailed finite-regime upper bound is given in 10.17 where $(k_0, 0)$ denotes the target and the value of the constant C_{p_0} (independent of K and ϵ) is

$$C_{p_0} = \frac{1}{\left(1 - \exp(-2(1 - 2(1 - p_0)^2)^2)\right)^2}. \tag{10.13}$$

Also, CBRW satisfies the reliability constraint:

$$\mathbb{P}[\mathcal{S}_\delta \neq \mathcal{S}] \leq \epsilon. \tag{10.14}$$

proof 24 *See Appendix for the proof.*

10.3.2 Hierarchical Target Setting

With a hierarchical target at level l_0, let $\{\mathcal{T}_{l_0+1}, ..., \mathcal{T}_L\}$ be the same as defined for a target that is a decedent of the hierarchical target. Also, define $\mathcal{T}_1'$ and $\mathcal{T}_2'$ as subtrees whose root nodes are children of the hierarchical target. Let (k_l', l) denote the root node of $\mathcal{T}_l'$.

Theorem 29 *Provided that the gap in the mean value of the processes and the threshold is bounded away from zero at all levels of the tree, the sample complexity of CBRW is in optimal logarithmic order in both K and $\frac{1}{\epsilon}$:*

$$\mathbb{E}_{\mathcal{F}}[T_{RW-CB}] = O(\log_2 K + \log \frac{1}{\epsilon}).$$

A detailed finite-regime upper bound is given in 10.18 where (k_{l_0}, l_0) denotes the hierarchical target and the value of the constant $C_{p_0}^H$ (independent of K and ϵ) is

$$C_{p_0}^H = \frac{1}{\left(1 - \exp(-2(1 - 2(1 - p_0)^3)^2)\right)^2}. \tag{10.15}$$

Also, CBRW satisfies the reliability constraint:

$$\mathbb{P}[\mathcal{S}_\delta \neq \mathcal{S}] \leq \epsilon. \tag{10.16}$$

proof 25 *See Appendix for the proof.*

10.4 Extensions

In this section we provide discussion on the extensions of CBRW and its analysis to multiple target detection, general reward distributions, general tree structures and different applications.

$$\mathbb{E}_{\mathcal{F}}[T_{CBRW}] \leq 2\sum_{l=1}^{L} C_{p_0}\Big(\frac{48}{(\mu_{k_l,l}-\eta_l)^2}\log\frac{24\sqrt[3]{\frac{2}{p_0}}}{(\mu_{k_l,l}-\eta_l)^2}+2\Big) + \frac{48}{(\mu_{k_0,0}-\eta_0)^2}\log\frac{24\sqrt[3]{\frac{2C_{p_0}L}{\epsilon}}}{(\mu_{k_0,0}-\eta_0)^2}+2. \quad (10.17)$$

$$\begin{aligned}\mathbb{E}_{\mathcal{F}}[T_{CBRW}] \leq\ & 3\sum_{l=1_0+1}^{L}\frac{C^H_{p_0}}{1-p_0}\Big[\Big(\frac{48}{(\mu_{k_l,l}-\eta_l)^2}\log\frac{24\sqrt[3]{\frac{2}{p_0}}}{(\mu_{k_l,l}-\eta_l)^2}+2\Big) + \frac{p_0}{(1-p_0)}\frac{16\log 2}{(\mu_{k_l,l}-\eta_l)^2}\Big] \\ & +3\sum_{l'=1}^{2}\frac{C^H_{p_0}}{1-p_0}\Big[\Big(\frac{48}{(\mu_{k'_l,l'}-\eta_{l'})^2}\log\frac{24\sqrt[3]{\frac{2}{p_0}}}{(\mu_{k'_l,l'}-\eta_{l'})^2}+2\Big) + \frac{p_0}{(1-p_0)}\frac{16\log 2}{(\mu_{k_{l'},l'}-\eta_{l'})^2}\Big] \\ & +\log_2\frac{6LC^H_{p_0}p_0}{\epsilon}\Big(\frac{48}{(\mu_{k_{l_0},l_0}-\eta_{l_0})^2}\log\frac{24\sqrt[3]{\frac{4}{\epsilon}}}{(\mu_{k_{l_0},l_0}-\eta_{l_0})^2}+2\Big). \quad (10.18)\end{aligned}$$

Figure 10.3: Finite-regime upper bounds on the performance of CBRW under leaf-level 10.17 and hierarchical 10.18 target settings.

10.4.1 Multiple Targets

Detecting $|\mathcal{S}| > 1$ targets can be easily implemented by sequentially locating the targets one by one. We assume that each target can be removed after it is located by CBRW.

Under leaf-level target setting, each leaf target is removed after being detected. For example, in noisy group testing, each defective item will be removed after being located and the search will be repeated for the detection of the next defective item.

Under hierarchical target setting, as well, each hierarchical target is removed

after being detected. For example, in HHH detection, a counter is assigned to each detected HHH and its count will be subtracted from all parents of the detected HHH.

Under both leaf-level and hierarchical target settings this consideration results in $O(|\mathcal{S}| \log K + |\mathcal{S}| \log \frac{1}{\epsilon})$ sample complexity.

10.4.2 Heavy-Tailed Distributions

The extension to more general distribution models can be implemented by only modifying the local CB-based test $\mathcal{A}$. The modification is implemented in a way that the confidence levels remain the same. As a result, the behavior of the random walk on the tree remains the same.

Specifically, for heavy-tailed distributions with existing b'th moment as given in 10.3, we modify the test $\mathcal{A}$ with

$$\widehat{X}(s, p_1) + 4u^{1/b}\left(\frac{\log \frac{2s^3}{p_1}}{s}\right)^{\frac{b-1}{b}} \quad < \quad \eta \textit{ replacing } 10.6, \textit{ and}$$

$$\widehat{X}(s, p_2) - 4u^{1/b}\left(\frac{\log \frac{2s^3}{p_2}}{s}\right)^{\frac{b-1}{b}} \quad > \quad \eta \textit{ replacing } 10.7.$$

The resulting CBRW achieves the same $O(\log K + \log \frac{1}{\epsilon})$ sample complexity under both leaf-level and hierarchical target settings. The proofs follow similar to the proofs of Theorems 28 and 29, and Lemma 9, using confidence bounds 10.5 instead of 10.2 in the proof of Lemma 9.

10.4.3 General Tree Structure

Consider a general (not necessarily binary) tree model for data aggregation as shown in Fig. 10.4. The CBRW policy can be extended to general tree structured model with following modifications.

To have the required confidence level in taking the steps toward the target, the input parameters in local CB-based test $\mathcal{A}$ are modified based on the degree $d_{k,l}$ of each node (k,l) in the tree. In particular, under leaf-level target setting, we choose $p_0 \in (1 - \frac{1}{2^{-(d_{k,l}-1)}})$ and $p_2 = \frac{\epsilon}{(D-1)LC}$ where L is the maximum distance from the root node to a leaf node, D is the maximum degree of the nodes in the tree and C is a constant independent of K and ϵ. Under the hierarchical target setting, we choose $p_0 \in (1 - \frac{1}{2^{-d_{k,l}}})$ and when increasing the confidence level iteratively to detect the hierarchical target, we terminate the search when p goes below $\frac{\epsilon}{(D-1)LC}$.

In the global random walk, the decision to move the pointer to a child or the parent of the current location is accordingly made based on the outputs of test $\mathcal{A}$. Following similar lines as in the proof of Theorems 28 and 29, we can show a sample complexity of $O(LD) + O(\log \frac{1}{\epsilon})$ under both leaf-level and hierarchical target settings.

10.4.4 Variations for Different Applications

The CBRW policy directly applies to leaf-level heavy hitter and HHH detection under leaf-level and hierarchical target settings, respectively. In particular, provided a controllable counter which can be assigned to each IP prefix, we assign

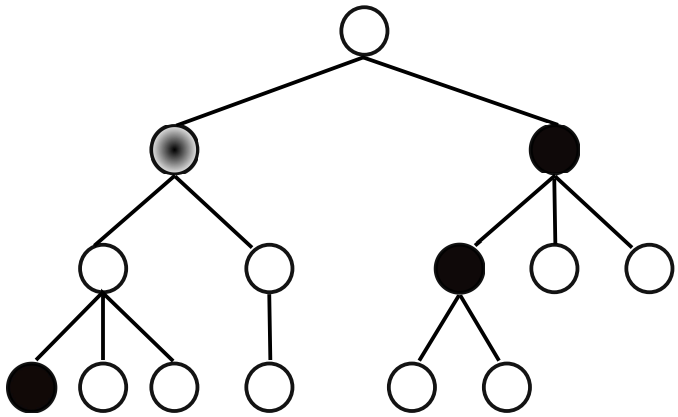

Figure 10.4: An example of a general (not necessarily binary) hierarchical data streams model.

the counter to the current location which is desired to be tested according to $\mathcal{A}$. Based on the outcome of the tests, the counter is moved on the tree according to CBRW. When there are several counters available the tree can be partitioned to smaller subtrees and CBRW is run on each subtree separately to make an efficient use of the available counters.

The CBRW policy under leaf-level target setting directly applies to noisy group testing where the defective items are the leaf-level targets

The extension of the result to active classification is not immediately clear. To formalize the problem, let the input space be the $[0, 1]$ interval. We limit the input space to be one-dimensional in order to demonstrate the main ideas that are relevant to our work. The hypothesis class denoted by $\mathcal{H}$, is the set of all step functions on [0,1] interval.

$$\mathcal{H} = \left\{ h_z : [0,1] \to \mathbb{R}, h_z(x) = \mathbb{1}_{\{(z,1]\}}(x), z \in (0,1) \right\} \tag{10.19}$$

Each hypothesis h_z assigns a binary label to each element of the input space

$[0, 1]$. There is a true hypothesis h_{z^*} that determines the ground truth labels for the input space.

The learner is allowed to make sequential observations by adaptively sampling h_{z^*}. The observations are however possibly noisy. The goal is to design a sequential sampling strategy aiming at minimizing the sample complexity required to obtain a confidence interval of length Δ for z^* at a $1 - \epsilon$ confidence level. Specifically, the learner chooses the sampling point x at each time t and receives a noisy sample of the true hypothesis. We consider two noise models with unknown distribution.

i) In the first noise model, the learner receives a noisy value of the threshold function in the form of:

$$h^N_{z^*}(x; t) = h_{z^*}(x; t) + N(x, t), \tag{10.20}$$

where $N(x, t)$ is a zero mean sampling noise that possibly depends on the sampling point x and is generated i.i.d. over t.

ii) In the second noise model, the binary samples can flip from zeros to ones and vice versa. Specifically, the learner receives erroneous binary samples with an error probability of $p(.)$ in the form of

$$h^B_{z^*}(x; t) = h_{z^*}(x; t) \oplus B(x, t), \tag{10.21}$$

where $\oplus$ is the boolean sum and $B(x, t)$ is a Bernoulli random variable with $\mathbb{P}[B(x, t) = 1] = p(x)$ that possibly depends on the sampling point x and is generated i.i.d. over t.

We now present a solution to the adaptive sampling problem based on the

results obtained for CBRW strategy. For the simplicity of presentation we assume $\Delta = \frac{1}{2^L}$ ($K = \frac{1}{\Delta}$). Let each node on a binary tree $\mathcal{T}$ represent an interval $[z^L_{k,l}, z^R_{k,l}] \subset [0,1]$ with $z^L_{k,l} = (k2^{L-l} - 1)\Delta$ and $z^U_{k,l} = k2^{L-l}\Delta$. Notice that the interval corresponding to each node is the union of the intervals corresponding to its decedents.

A test $\mathcal{A}'$ (similar to $\mathcal{A}$) is considered which upon calling on a node on $\mathcal{T}$ determines whether z^* is likely to be in the corresponding interval or not (outputs 1 or 0, respectively). In particular, $\mathcal{A}'$ consists of calling $\mathcal{A}$ on the random samples from both boundary points of each interval. The output of $\mathcal{A}'$ is 1 (indicating that the interval is likely to contain z^*) if and only if the output of $\mathcal{A}$ is 0 on the left boundary and 1 on the right boundary. The output of $\mathcal{A}'$ is 0 otherwise. The parameters are chosen as $p_1 = p_2 = p$ with $p \in (0, 1 - \frac{1}{\sqrt[4]{2}})$ on a non-leaf node. When at a leaf node choose $p_2 = \frac{\epsilon}{2LC_p}$ to examine the interval for containing z^*. On all levels of tree $\eta_l = 0.5$. From the results on the analysis of CBRW the above solution has a sample complexity of $O(\frac{1}{c^2}\log K + \frac{1}{c^2}\log\frac{1}{\epsilon})$ where c is 0.5 under the first noise model and c is a lower bound on the gap in $0.5 - p(.)$ in the second noise model.

10.5 Conclusion

In this paper, we studied the problem of detecting a few targets among a large number of hierarchical data streams modeled as random processes with unknown distributions. We designed sequential strategies to interactively choose to observations point aiming at minimizing the sample complexity. The proposed strategies detect the targets at the desired confidence level with an order

optimal logarithmic sample complexity in both problem size and the parameter of reliability constraint. We further showed the extensions of the results to a number of active inference and learning problems in networking and data analytics applications.

APPENDIX A

PROOFS OF LEMMAS AND THEOREMS FROM PART I

A.1 Proof of Lemma 2

Let $\overline{\mu}_s$ be the sample mean obtained from s i.i.d. observations. By Chernoff-Hoeffding bound 3.3, for all $a \in (0, \frac{1}{2\zeta_0}]$,

$$\begin{cases} \mathbb{P}[\overline{\mu}_s - \mu(X) < -\delta_1] & \leq \exp(-as\delta_1^2), \\ \mathbb{P}[\overline{\mu}_s - \mu(X) > \delta_1] & \leq \exp(-as\delta_1^2), \end{cases}$$

and, for all $a \in (0, \frac{1}{2\zeta_1}]$,

$$\begin{cases} \mathbb{P}[|\frac{1}{s}\sum_{t=1}^{s}(X(t) - \mu(X))^2 - \sigma^2(X)| < -\delta_2] & \leq \exp(-as\delta_2^2), \\ \mathbb{P}[|\frac{1}{s}\sum_{t=1}^{s}(X(t) - \mu(X))^2 - \sigma^2(X)| > \delta_2] & \leq \exp(-as\delta_2^2), \end{cases}$$

where $X(t)$ is the tth observation of the random variable X. The mean-variance deviation term can be written as

$$\begin{aligned} \overline{\xi}_s - \xi(X) &= \frac{1}{s}\sum_{t=1}^{s}(X(t) - \overline{\mu}_s)^2 - \rho\overline{\mu}_s - \xi(X) \\ &= \frac{1}{s}\sum_{t=1}^{s}(X(t) - \mu(X))^2 + (\mu(X) - \overline{\mu}_s)^2 \\ &+ \frac{2}{s}\sum_{t=1}^{s}(X(t) - \mu(X))(\mu(X) - \overline{\mu}_s) - \rho\overline{\mu}_s - \xi(X) \\ &= \frac{1}{s}\sum_{t=1}^{s}(X(t) - \mu(X))^2 - \sigma^2(X) - (\mu(X) - \overline{\mu}_s)^2 - \rho(\overline{\mu}_s - \mu(X)). \end{aligned} \quad \text{(A.1)}$$

Notice that the second term on the right hand side of A.1 is always negative. For $\delta_1 = \frac{\delta}{1+\rho}$ and $a \leq \frac{1}{2\zeta}$, substituting $\overline{\xi}_s - \xi(X)$ from A.1

$$\mathbb{P}[\overline{\xi}_s - \xi(X) > \delta] \leq \mathbb{P}[\frac{1}{s}\sum_{t=1}^{s}(X(t) - \mu(X))^2 - \sigma^2(X) > \delta_1] + \mathbb{P}[\overline{\mu}_s - \mu(X) < -\delta_1]$$

$$\begin{aligned} &\leq \exp(-as\delta_1^2) + \exp(-as\delta_1^2) \\ &= 2\exp(-\frac{as\delta^2}{(1+\rho)^2}). \end{aligned} \tag{A.2}$$

To prove 5.7 let $\delta_1 = \frac{\delta}{2+\rho}$. Notice that, $\delta_1 \leq 1$ when $\delta \leq 2+\rho$ and $(\mu(X) - \overline{\mu}_s) < \delta_1$ implies $(\mu(X) - \overline{\mu}_s)^2 < \delta_1$. For $a \leq \frac{1}{2\zeta}$, substituting $\overline{\xi}_s - \xi(X)$ from A.1

$$\begin{aligned} \mathbb{P}[\overline{\xi}_s - \xi(X) < -\delta] &\leq \mathbb{P}[\frac{1}{s}\sum_{t=1}^{s}(X(t) - \mu(X))^2 - \sigma^2(X) < -\delta_1] + \mathbb{P}[\overline{\mu}_s - \mu(X) > \delta_1] \\ &\leq \exp(-as\delta_1^2) + \exp(-as\delta_1^2) \\ &= 2\exp(-\frac{as\delta^2}{(2+\rho)^2}). \quad \blacksquare \end{aligned} \tag{A.3}$$

A.2 Proof of Lemma 3

Let $\overline{\mu}_\pi = \frac{1}{T}\sum_{t=1}^{T} X_{\pi(t)}(t)$ and $\mu_\pi = \mathbb{E}[\overline{\mu}_\pi]$. In order to show the expression of regret given in 3, we expand the cumulative variance term.

$$\begin{aligned} \mathbb{E}[\sum_{t=1}^{T}(X_{\pi(t)}(t) - \overline{\mu}_\pi)^2] &= \mathbb{E}[\sum_{i=1}^{K}\sum_{s=1}^{\tau_i(T)}(X_i(t_i(s)) - \overline{\mu}_\pi)^2] \\ &= \mathbb{E}[\sum_{i=1}^{K}\sum_{s=1}^{\tau_i(T)}(X_i(t_i(s)) - \mu_i + \mu_i - \overline{\mu}_\pi)^2] \\ &= \mathbb{E}[\sum_{i=1}^{K}\sum_{s=1}^{\tau_i(T)}((X_i(t_i(s)) - \mu_i)^2 + (\mu_i - \overline{\mu}_\pi)^2 \\ &\quad + 2(X_i(s) - \mu_i)(\mu_i - \overline{\mu}_\pi))]. \end{aligned} \tag{A.4}$$

The first term on the RHS of A.4 equals to, by Wald identity,

$$\mathbb{E}[\sum_{i=1}^{K}\sum_{s=1}^{\tau_i(T)}((X_i(s) - \mu_i)^2] = \sum_{i=1}^{K}\mathbb{E}\tau_i\sigma_i^2. \tag{A.5}$$

The second term can be written as

$$\mathbb{E}[\sum_{i=1}^{K}\tau_i(\mu_i - \overline{\mu}_\pi)^2] = \mathbb{E}[\sum_{i=1}^{K}\tau_i(\mu_i - \mu_\pi + \mu_\pi - \overline{\mu}_\pi)^2]$$

$$= \mathbb{E}[\sum_{i=1}^{K} \tau_i((\mu_i - \mu_\pi)^2 + (\mu_\pi - \overline{\mu}_\pi)^2 + 2(\mu_i - \mu_\pi)(\mu_\pi - \overline{\mu}_\pi))]$$

$$= \sum_{i=1}^{K} \mathbb{E}[\tau_i(\mu_i - \mu_\pi)^2] + \mathbb{E}[T(\mu_\pi - \overline{\mu}_\pi)^2] + 2\mathbb{E}[\sum_{i=1}^{K} \tau_i(\mu_i - \mu_\pi)(\mu_\pi - \overline{\mu}_\pi)]. \quad (A.6)$$

The third term can be written as

$$\mathbb{E}[\sum_{i=1}^{K} 2\tau_i(\overline{\mu}_i - \mu_i)(\mu_i - \overline{\mu}_\pi)] = 2\mathbb{E}[\sum_{i=1}^{K} \tau_i(\overline{\mu}_i - \mu_i)(\mu_i - \mu_\pi) + \tau_i(\overline{\mu}_i - \mu_i)(\mu_\pi - \overline{\mu}_\pi)]$$

$$= 2\mathbb{E}[\sum_{i=1}^{K} \tau_i(\overline{\mu}_i - \mu_i)(\mu_\pi - \overline{\mu}_\pi)]. \quad (A.7)$$

From A.5, A.6 and A.7, we have

$$\mathbb{E}[\sum_{t=1}^{T}(X_{\pi(t)}(t) - \overline{\mu}_\pi)^2] = \sum_{i=1}^{K} \mathbb{E}\tau_i\sigma_i^2 + \sum_{i=1}^{K} \mathbb{E}[\tau_i(\mu_i - \mu_\pi)^2] + \mathbb{E}[T(\mu_\pi - \overline{\mu}_\pi)^2] + 2\mathbb{E}[\sum_{i=1}^{K} \tau_i(\mu_i - \mu_\pi)(\mu_\pi - \overline{\mu}_\pi)] + 2\mathbb{E}[\sum_{i=1}^{K} \tau_i(\overline{\mu}_i - \mu_i)(\mu_\pi - \overline{\mu}_\pi)]$$

$$= \sum_{i=1}^{K} \mathbb{E}\tau_i\sigma_i^2 + \sum_{i=1}^{K} \mathbb{E}[\tau_i(\mu_i - \mu_\pi)^2] + \mathbb{E}[T(\mu_\pi - \overline{\mu}_\pi)^2] + 2\mathbb{E}[\sum_{i=1}^{K} \tau_i(\overline{\mu}_i - \mu_\pi)(\mu_\pi - \overline{\mu}_\pi)]$$

$$= \sum_{i=1}^{K} \mathbb{E}\tau_i\sigma_i^2 + \sum_{i=1}^{K} \mathbb{E}[\tau_i(\mu_i - \mu_\pi)^2] + \mathbb{E}[T(\mu_\pi - \overline{\mu}_\pi)^2] - 2\mathbb{E}[T(\mu_\pi - \overline{\mu}_\pi)^2] \quad (A.8)$$

$$= \sum_{i=1}^{K} \mathbb{E}\tau_i\sigma_i^2 + \sum_{i=1}^{K} \mathbb{E}[\tau_i(\mu_i - \mu_\pi)^2] - \mathbb{E}[T(\mu_\pi - \overline{\mu}_\pi)^2]. \quad (A.9)$$

To arrive at A.8, $\sum_{i=1}^{K} \tau_i(T) = T$ and $\sum_{i=1}^{K} \tau_i(T)\overline{\mu}_i(T) = T\overline{\mu}_\pi$ are used. Similarly, $\sum_{i=1}^{K} \tau_i(T)\mu_i = T\mu_\pi$ and it can be shown that

$$\sum_{i=1}^{K} \mathbb{E}[\tau_i(\mu_i - \mu_\pi)^2] = \sum_{i=1}^{K} \mathbb{E}[\tau_i(\mu_i - \frac{\sum_{j=1}^{K} \mathbb{E}\tau_j\mu_j}{T})^2]$$

$$
\begin{aligned}
&= \frac{1}{T^2}\sum_{i=1}^{K}\mathbb{E}[\tau_i(\sum_{j=1}^{K}\mathbb{E}\tau_j\Gamma_{i,j})^2] \\
&= \frac{1}{T^2}\mathbb{E}[\tau_*(\sum_{j=1}^{K}\mathbb{E}\tau_j\Gamma_{*,j})^2] + \frac{1}{T^2}\sum_{i\neq *}\mathbb{E}[\tau_i(\sum_{j\neq *}\mathbb{E}\tau_j\Gamma_{i,j} + \mathbb{E}\tau_*\Gamma_{i,*})^2] \\
&= \frac{1}{T^2}\mathbb{E}[(T-\sum_{i\neq *}\tau_i)(\sum_{j=1}^{K}\mathbb{E}\tau_j\Gamma_{*,j})^2] + \frac{1}{T^2}\sum_{i\neq *}\mathbb{E}[\tau_i(\sum_{j\neq *}\mathbb{E}\tau_j\Gamma_{i,j} + (T-\sum_{j\neq *}\mathbb{E}\tau_j)\Gamma_{i,*})^2] \\
&= \frac{1}{T}(\sum_{j=1}^{K}\mathbb{E}\tau_j\Gamma_{*,j})^2 - \frac{1}{T^2}\sum_{i\neq *}\mathbb{E}[\tau_i(\sum_{j=1}^{K}\mathbb{E}\tau_j\Gamma_{*,j})^2] \\
&\quad + \frac{1}{T^2}\sum_{i\neq *}\mathbb{E}[\tau_i(T\Gamma_{i,*} + \sum_{j\neq *}\mathbb{E}\tau_j\Gamma_{*,j})^2] \\
&= \frac{1}{T}(\sum_{j=1}^{K}\mathbb{E}\tau_j\Gamma_{*,j})^2 - \frac{1}{T^2}\sum_{i\neq *}\mathbb{E}[\tau_i(\sum_{j=1}^{K}\mathbb{E}\tau_j\Gamma_{*,j})^2] \\
&\quad + \sum_{i\neq *}\mathbb{E}\tau_i\Gamma_{i,*}^2 + \frac{1}{T^2}\sum_{i\neq *}\mathbb{E}[\tau_i(\sum_{j\neq *}\mathbb{E}\tau_j\Gamma_{*,j})^2] + \frac{2}{T}\sum_{i\neq *}\mathbb{E}[\tau_i\Gamma_{i,*}(\sum_{j=1}^{K}\mathbb{E}\tau_j\Gamma_{*,j})] \\
&= \sum_{i\neq *}\mathbb{E}\tau_i\Gamma_{i,*}^2 - \frac{1}{T}(\sum_{j=1}^{K}\mathbb{E}\tau_j\Gamma_{*,j})^2. \qquad \text{(A.10)}
\end{aligned}
$$

For the third term on the RHS of A.9, we have

$$
\begin{aligned}
&\mathbb{E}[T(\mu_\pi - \overline{\mu}_\pi)^2] \\
&= \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K}\sum_{s=1}^{\tau_i}X_i(t_i(s)) - \sum_{i=1}^{K}\mathbb{E}\tau_i\mu_i)^2] \\
&= \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K}\sum_{s=1}^{\tau_i}(X_i(t_i(s)) - \mu_i) + \sum_{i=1}^{K}(\tau_i - \mathbb{E}\tau_i)\mu_i)^2] \\
&= \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K}\sum_{s=1}^{\tau_i}(X_i(t_i(s)) - \mu_i))^2] + \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K}(\tau_i - \mathbb{E}\tau_i)\mu_i)^2] \\
&\quad + \frac{2}{T}\mathbb{E}[(\sum_{i=1}^{K}\sum_{s=1}^{\tau_i}(X_i(t_i(s)) - \mu_i))(\sum_{i=1}^{K}(\tau_i - \mathbb{E}\tau_i)\mu_i)] \\
&= \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K}\tau_i(\overline{\mu}_i - \mu_i))^2] + \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K}(\tau_i - \mathbb{E}\tau_i)\Gamma_{i,*})^2] \\
&\quad + \frac{2}{T}\mathbb{E}[(\sum_{i=1}^{K}\tau_i(\overline{\mu}_i - \mu_i))(\sum_{i=1}^{K}(\tau_i - \mathbb{E}\tau_i)\Gamma_{i,*})] \qquad \text{(A.11)}
\end{aligned}
$$

$$= \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K}\tau_i(\overline{\mu}_i-\mu_i))^2]+\frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K}\tau_i\Gamma_{i,*})^2]-\frac{1}{T}(\sum_{i=1}^{K}\mathbb{E}\tau_i\Gamma_{i,*})^2$$
$$+\frac{2}{T}\mathbb{E}[(\sum_{i=1}^{K}\tau_i(\overline{\mu}_i-\mu_i))(\sum_{i=1}^{K}\tau_i\Gamma_{i,*})]. \quad (A.12)$$

Equation A.11 follows from $\sum_{i=1}^{K}(\tau_i-\mathbb{E}\tau_i)\mu_i = (\tau_*-\mathbb{E}\tau_*)\mu_* + \sum_{i\neq *}(\tau_i-\mathbb{E}\tau_i)\mu_i = -\sum_{i\neq *}(\tau_i-\mathbb{E}\tau_i)\mu_* + \sum_{i\neq *}(\tau_i-\mathbb{E}\tau_i)\mu_i = \sum_{i\neq *}(\tau_i-\mathbb{E}\tau_i)\Gamma_{i,*}$. We know that for any random variable X, $\sigma^2(X) = \mathbb{E}[X^2]-\mathbb{E}[X]^2$. To arrive at A.12, set $X = \sum_{i=1}^{K}\tau_i\Gamma_{i,*}$ also notice that $\frac{2}{T}\mathbb{E}[(\sum_{i=1}^{K}\tau_i(\overline{\mu}_i-\mu_i)(\sum_{i=1}^{K}\mathbb{E}\tau_i\Gamma_{i,*})] = 0$.

Thus, from A.9, A.10 and A.12, we have

$$\mathbb{E}[\sum_{t=1}^{T}(X_{\pi(t)}(t)-\overline{\mu}_\pi)^2] = \sum_{i=1}^{K}\mathbb{E}\tau_i\sigma_i^2+\sum_{i\neq *}\mathbb{E}\tau_i\Gamma_{i,*}^2-\frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K}\tau_i(\overline{\mu}_i-\mu_i))^2]-\frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K}\tau_i\Gamma_{i,*})^2]$$
$$-\frac{2}{T}\mathbb{E}[\sum_{i=1}^{K}\tau_i(\overline{\mu}_i-\mu_i)(\sum_{i=1}^{K}\tau_i\Gamma_{i,*})]$$
$$= \sum_{i=1}^{K}\mathbb{E}\tau_i\sigma_i^2+\sum_{i\neq *}\mathbb{E}\tau_i\Gamma_{i,*}^2-\frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K}\tau_i(\overline{\mu}_i-\mu_*))^2]. \quad (A.13)$$

Now we can show the expression for $\widehat{R}_\pi(T)$ for any policy π that plays arm i for τ_i times

$$\widehat{R}_\pi(T) = \xi_\pi(T)-\xi_{\widehat{\pi}^*}(T)$$
$$= \sum_{i=1}^{K}\mathbb{E}\tau_i\xi_i+\sum_{i\neq *}\mathbb{E}\tau_i\Gamma_{i,*}^2-\frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K}\tau_i(\overline{\mu}_i-\mu_*))^2]-T\xi_*+\frac{1}{T}\mathbb{E}[T(\overline{\mu}_*-\mu_*)]$$
$$= \sum_{i=1}^{K}\mathbb{E}\tau_i\Delta_i+\sum_{i=1}^{K}\mathbb{E}\tau_i\Gamma_{i,*}^2-\frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K}\tau_i(\overline{\mu}_i-\mu_*))^2]+\sigma_*^2$$

as desired.

A.3 Proof of Lemma 4

In order to prove 5.11, we write the expected value of $\tau(\overline{\mu}-\mu)^2$ divided by $\log T$ as integrating the tail probability. For the tail probability we have, for a real number $x > 0$

$$\begin{aligned}
\mathbb{P}[\tau(\overline{\mu}-\mu)^2 > x\log T] &\leq \mathbb{P}[\max_{1\leq s\leq T} s(\overline{\mu}_s-\mu)^2 > x\log T] \\
&= \mathbb{P}[\max_{1\leq s\leq T} \sqrt{s}|\overline{\mu}_s-\mu| > \sqrt{x\log T}] \\
&\leq \sum_{s=1}^{T} \mathbb{P}[|\overline{\mu}_s-\mu| > \sqrt{\frac{x\log T}{s}}] \\
&\leq \sum_{s=1}^{T} 2\exp(-ax\log T) \\
&= 2T^{-ax+1}.
\end{aligned}$$

Now, we can write

$$\begin{aligned}
\mathbb{E}[\frac{\tau(\overline{\mu}-\mu)^2}{\log T}] &= \int_0^{\infty} \mathbb{P}[\frac{\tau(\overline{\mu}-\mu)^2}{\log T} > x]dx \\
&\leq \frac{1}{a} + \int_{\frac{1}{a}}^{\infty} \mathbb{P}[\frac{\tau(\overline{\mu}-\mu)^2}{\log T} > x]dx \\
&\leq \frac{1}{a} + \int_{\frac{1}{a}}^{\infty} 2T^{-ax+1}dx \\
&= \frac{1}{a} + 2\frac{T^{-ax+1}}{a\log T}|_{\infty}^{\frac{1}{a}} \\
&= \frac{1}{a}(1+\frac{2}{\log T}).
\end{aligned}$$

Thus, multiplying by $\log T$, we have

$$\mathbb{E}[\tau(\overline{\mu}-\mu)^2] \leq \frac{1}{a}(\log T + 2).$$

A.4 Proof of Theorem 12

Since $\xi_{\pi^*} \leq \xi_{\widehat{\pi}^*}$, it is straightforward to see that

$$R_\pi(T) - \widehat{R}_\pi(T) = \xi_\pi(T) - \xi_{\pi^*}(T) - (\xi_\pi(T) - \xi_{\widehat{\pi}^*}(T)) \geq 0. \tag{A.14}$$

For the upper bound, we have

$$\begin{aligned} R_\pi(T) - \widehat{R}_\pi(T) &= \xi_{\widehat{\pi}^*}(T) - \xi_{\pi^*}(T) \\ &= -\widehat{R}_{\pi^*}(T). \end{aligned} \tag{A.15}$$

From Lemma 3, we have

$$\widehat{R}_{\pi^*}(T) = \sum_{i=1}^{K} \mathbb{E}\tau_i \Delta_i + \sum_{i=1}^{K} \mathbb{E}\tau_i \Gamma_{i,*}^2 - \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K} \tau_i(\overline{\mu}_i - \mu_*))^2] + \sigma_*^2, \tag{A.16}$$

where τ_i are the number of times arm i is played by π^*. We have, by Cauchy-Schwartz inequality,

$$\begin{aligned} \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K} \tau_i(\overline{\mu}_i - \mu_*))^2] &= \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K} \tau_i(\overline{\mu}_i - \mu_i))^2] + \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K} \tau_i \Gamma_{i,*})^2] \\ &\quad + \frac{2}{T}\mathbb{E}[\sum_{i=1}^{K} \tau_i(\overline{\mu}_i - \mu_i)(\sum_{i=1}^{K} \tau_i \Gamma_{i,*})] \\ &\leq \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K} \tau_i(\overline{\mu}_i - \mu_i))^2] + \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K} \tau_i \Gamma_{i,*})^2] \\ &\quad + \frac{2}{T}\sqrt{\mathbb{E}[(\sum_{i=1}^{K} \tau_i(\overline{\mu}_i - \mu_i))^2]\mathbb{E}[(\sum_{i=1}^{K} \tau_i \Gamma_{i,*})^2]} \\ &= \frac{1}{T}\sum_{i=1}^{K} \mathbb{E}\tau_i \sigma_i^2 + \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K} \tau_i \Gamma_{i,*})^2] \\ &\quad + \frac{2}{T}\sqrt{\sum_{i=1}^{K} \mathbb{E}\tau_i \sigma_i^2}\sqrt{\mathbb{E}[(\sum_{i=1}^{K} \tau_i \Gamma_{i,*})^2]}. \end{aligned} \tag{A.17}$$

To arrive at A.17, we also use $\mathbb{E}[(\sum_{i=1}^{K}\tau_i(\overline{\mu}_i-\mu_i))^2] = \sum_{i=1}^{K}\mathbb{E}\tau_i\sigma_i^2$ as a result of Wald's second identity. For the second term on the RHS we have, by applying again Cauchy-Schwartz inequality,

$$\begin{aligned}\frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K}\tau_i\Gamma_{i,*})^2] &\leq \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K}\tau_i)(\sum_{i=1}^{K}\tau_i\Gamma_{i,*}^2)] \\ &= \sum_{i=1}^{K}\mathbb{E}\tau_i\Gamma_{i,*}^2. \end{aligned} \tag{A.18}$$

For a set of positive real numbers h_i, we have $\sqrt{\sum_i h_i} \leq \sum_i \sqrt{h_i}$. We can apply this inequality to the third term on the RHS of A.17 and from A.17, A.18 and $\frac{1}{T}\sum_{i=1}^{K}\mathbb{E}\tau_i\sigma_i^2 \leq \sigma_{\max}^2$ (where $\sigma_{\max} = \max_i \sigma_i$) , we have

$$\frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K}\tau_i(\overline{\mu}_i-\mu_*))^2] \leq \sigma_{\max}^2 + \sum_{i=1}^{K}\mathbb{E}\tau_i\Gamma_{i,*}^2 + 2\sigma_{\max}\sum_{i=1}^{K}\sqrt{\mathbb{E}\tau_i\Gamma_{i,*}^2}. \tag{A.19}$$

Thus we can write

$$\begin{aligned}\widehat{R}_{\pi^*}(T) &\geq \sum_{i=1}^{K}\mathbb{E}\tau_i\Delta_i + \sum_{i=1}^{K}\mathbb{E}\tau_i\Gamma_{i,*}^2 - \sum_{i=1}^{K}\mathbb{E}\tau_i\Gamma_{i,*}^2 - 2\sigma_{\max}\sqrt{\sum_{i=1}^{K}\mathbb{E}\tau_i\Gamma_{i,*}^2} + \sigma_*^2 - \sigma_{\max}^2 \\ &\geq \sum_{i=1}^{K}\mathbb{E}\tau_i\Delta_i - 2\sigma_{\max}\sum_{i=1}^{K}\sqrt{\mathbb{E}\tau_i}|\Gamma_{i,*}| - \sigma_{\max}^2 \\ &\geq \sum_{i=1}^{K}\min_{x\geq 0}(x^2\Delta_i - 2\sigma_{\max}|\Gamma_{i,*}|x) - \sigma_{\max}^2 \\ &= -\sum_{i\neq *}\frac{\sigma_{\max}^2\Gamma_{i,*}^2}{\Delta_i} - \sigma_{\max}^2. \end{aligned}$$

This gives a lower bound on $\widehat{R}_{\pi^*}(T)$ which translates to an upper bound on $R_\pi(T) - \widehat{R}_\pi(T)$ by A.15. Although this lower bound is a constant independent of T, it grows unboundedly when Δ approaches 0. We next drive another lower bound on $\widehat{R}_{\pi^*}(T)$ that is independent of Δ.

$$\widehat{R}_\pi(T) = \sum_{i=1}^{K}\mathbb{E}\tau_i\Delta_i + \sum_{i\neq *}\mathbb{E}\tau_i\Gamma_{i,*}^2 - \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K}\tau_i(\overline{\mu}_i-\mu_*)^2] + \sigma_*^2$$

$$
\begin{aligned}
&\geq \sum_{i=1}^{K} \mathbb{E}\tau_i\Delta_i + \sum_{i\neq *} \mathbb{E}\tau_i\Gamma_{i,*}^2 - \sum_{i=1}^{K} \mathbb{E}[\tau_i(\overline{\mu}_i - \mu_*)^2] && \text{(A.20)}\\
&= \sum_{i=1}^{K} \mathbb{E}\tau_i\Delta_i + \sum_{i\neq *} \mathbb{E}\tau_i\Gamma_{i,*}^2 - \sum_{i=1}^{K} \mathbb{E}[\tau_i(\overline{\mu}_i - \mu_i + \Gamma_{i,*})^2]\\
&= \sum_{i=1}^{K} \mathbb{E}\tau_i\Delta_i - \sum_{i=1}^{K} \mathbb{E}[\tau_i(\overline{\mu}_i - \mu_i)^2]\\
&\geq \sum_{i=1}^{K} \mathbb{E}\tau_i\Delta_i - \frac{K}{a}\log T && \text{(A.21)}\\
&\geq -\frac{K}{a}\log T.
\end{aligned}
$$

Inequality A.20 holds as a result of Cauchy-Schwartz inequality (similar to A.18) and A.21 holds by Lemma 4.

A.5 Proof of Lemma 5

For $k \neq *$, construct $\mathcal{F}^k$, by only changing the distribution of arm k to f_k' such that arm k is the optimal arm ($-\delta < \xi_k' - \xi_* < 0$) and $|I(f_k, f_*) - I(f_k, f_k')| \leq \epsilon$ for arbitrary small ϵ. The possibility of such a model is a result of Assumption 1. Let γ denote the log-likelihood ratio between the $\mathcal{F}$ and $\mathcal{F}^k$: $\gamma = \log \frac{f_k(X_k(t_k(1)))\ldots f_k(X_k(t_k(\tau_k)))}{f_k'(X_k(t_k(1)))\ldots f_k'(X_k(t_k(\tau_k)))}$. We show that it is unlikely to have $\tau_k < \frac{c_1 \log T}{I(f_k, f_k')}$ under two different scenarios for γ.

First, consider $\gamma > c_5 \log T$ for a constant $c_5 > c_1$. We have

$$
\begin{aligned}
\mathbb{P}_{\mathcal{F}}[\tau_k < \frac{c_1 \log T}{I(f_k, f_k')}, \gamma > c_5 \log T] &= \mathbb{P}_{\mathcal{F}}[\tau_k < \frac{c_1 \log T}{I(f_k, f_k')}, \sum_{s=1}^{\tau_k} \log \frac{f_k(X_k(s))}{f_k'(X_k(s))} > c_5 \log T]\\
&\leq \mathbb{P}_{\mathcal{F}}[\max_{t \leq \frac{c_1 \log T}{I(f_k, f_k')}} \sum_{s=1}^{t} \log \frac{f_k(X_k(s))}{f_k'(X_k(s))} > \frac{c_5 \log T}{I(f_k, f_k')} I(f_k, f_k')]\\
&\leq \mathbb{P}_{\mathcal{F}}[\max_{t \leq \frac{c_1 \log T}{I(f_k, f_k')}} \frac{I(f_k, f_k')}{c_1 \log T} \sum_{s=1}^{t} \log \frac{f_k(X_k(s))}{f_k'(X_k(s))} > \frac{c_5}{c_1} I(f_k, f_k')].
\end{aligned}
$$

By strong law of large numbers $\sum_{s=1}^{t} \log \frac{f_k(X_k(s))}{f_k'(X_k(s))} \to I(f_k, f_k')$ *a.s.* as $t \to \infty$. Notice

that $\mathbb{E}_{\mathcal{F}}[\log \frac{f_k(X_k(s))}{f'_k(X_k(s))}] = I(f_k, f'_k)$. Thus,
$\max_{t \le \frac{c_1 \log T}{I(f_k, f'_k)}} \frac{I(f_k, f'_k)}{c_1 \log T} \sum_{s=1}^{t} \log \frac{f_k(X_k(s))}{f'_k(X_k(s))} \to I(f_k, f'_k)$ *a.s.* as $T \to \infty$. So, we have

$$\mathbb{P}_{\mathcal{F}}[\max_{t \le \frac{c_1 \log T}{I(f_k, f'_k)}} \frac{I(f_k, f'_k)}{c_1 \log T} \sum_{s=1}^{t} \log \frac{f_k(X_k(s))}{f'_k(X_k(s))} > \frac{c_5}{c_1} I(f_k, f'_k)] \to 0, \ as \ T \to \infty.$$

So when $\gamma > c_5 \log T$, by strong law of large numbers, we have

$$\mathbb{P}_{\mathcal{F}}[\tau_k < \frac{c_1 \log T}{I(f_k, f'_k)}, \gamma > c_5 \log T] \to 0, \ as \ T \to \infty. \tag{A.22}$$

Also, when Assumption 2 is satisfied, $\log f'_k(X)$ and $\log f_k(X)$ have sub-Gaussian distributions. Thus, $\log \frac{f'_k(X)}{f_k(X)}$ has sub-Gaussian distribution and using Chernoff-Hoeffding bound we can prove an upper bound for $\mathbb{P}_{\mathcal{F}}[\tau_k < \frac{c_1 \log T}{I(f_k, f'_k)}, \gamma > c_5 \log T]$, for finite T. Specifically,

$$\begin{aligned}
\mathbb{P}_{\mathcal{F}}[\tau_k < \frac{c_1 \log T}{I(f_k, f'_k)}, \gamma > c_5 \log T] &= \mathbb{P}_{\mathcal{F}}[\tau_k < \frac{c_1 \log T}{I(f_k, f'_k)}, \sum_{s=1}^{\tau_k} \log \frac{f_k(X_k(s))}{f'_k(X_k(s))} > c_5 \log T] \\
&\le \mathbb{P}_{\mathcal{F}}[\max_{t < \frac{c_1 \log T}{I(f_k, f'_k)}} \sum_{s=1}^{t} \log \frac{f_k(X_k(s))}{f'_k(X_k(s))} > c_5 \log T] \\
&\le \sum_{t=1}^{\frac{c_1 \log T}{I(f_k, f'_k)}} \mathbb{P}_{\mathcal{F}}[\frac{1}{t} \sum_{s=1}^{t} \log \frac{f_k(X_k(s))}{f'_k(X_k(s))} - \frac{1}{t} c_1 \log T > \frac{1}{t} c_5 \log T - \frac{1}{t} c_1 \log T] \\
&\le \sum_{t=1}^{\frac{c_1 \log T}{I(f_k, f'_k)}} \mathbb{P}_{\mathcal{F}}[\frac{1}{t} \sum_{s=1}^{t} \log \frac{f_k(X_k(s))}{f'_k(X_k(s))} - I(f_k, f'_k) > \frac{1}{t} c_5 \log T - \frac{1}{t} c_1 \log T] \quad \text{(A.23)} \\
&\le \sum_{t=1}^{\frac{c_1 \log T}{I(f_k, f'_k)}} \exp(-a_1 (c_5 - c_1)^2 \log^2 T / t) \quad \text{(A.24)} \\
&\le \frac{c_1 \log T}{I(f_k, f'_k)} T^{-a_1 I(f_k, f'_k) \frac{(c_5 - c_1)^2}{c_1}}. \quad \text{(A.25)}
\end{aligned}$$

Inequality A.23 holds since $I(f_k, f'_k) \le \frac{1}{t} c_1 \log T$ and A.24 holds according to Chernoff-Hoeffding bound. We point out that the Chernoff-Hoeffding bound constant a_1 is different from the constant a since we have a different random variable here.

Next, we consider $\gamma \leq c_5 \log T$. By Markov inequality, we have

$$\begin{aligned}
\mathbb{P}_{\mathcal{F}^k}[\tau_k < \frac{c_1 \log T}{I(f_k, f_k')}] &= \mathbb{P}_{\mathcal{F}^k}[T - \tau_k \geq T - \frac{c_1 \log T}{I(f_k, f_k')}] \\
&\leq \frac{\mathbb{E}_{\mathcal{F}^k}[T - \tau_k]}{T - \frac{c_1 \log T}{I(f_k, f_k')}}.
\end{aligned} \tag{A.26}$$

We can change the probability measure from $\mathcal{F}$ to $\mathcal{F}^k$ as follows. Let $\mathcal{S}(T)$ be the set of all observations over a time horizon with length T that satisfy a particular event. We have

$$\begin{aligned}
\mathbb{P}_{\mathcal{F}}[\mathcal{S}(T)] &= \mathbb{E}_{\mathcal{F}}[\mathbb{I}_{\mathcal{S}(T)}] \\
&= \int_{\mathcal{S}(T)} \prod_{i=1}^{K} \prod_{s=1}^{\tau_i} f_i(x_i(t_i(s))) \prod_{i=1}^{K} \prod_{s=1}^{\tau_i} dx_i(t_i(s)) \\
&= \int_{\mathcal{S}(T)} \prod_{i=1, i\neq k}^{K} \prod_{s=1}^{\tau_i} f_i(x_i(t_i(s))) \prod_{s=1}^{\tau_k} f_k'(x_i(t_i(s))) \frac{f_k(x_i(t_i(s)))}{f_k'(x_i(t_i(s)))} \prod_{i=1}^{K} \prod_{s=1}^{\tau_i} dx_i(t_i(s)) \\
&= \mathbb{E}_{\mathcal{F}^k}[\mathbb{I}_{\mathcal{S}(T)} \Pi_{s=1}^{\tau_k} \frac{f_k(x_i(t_i(s)))}{f_k'(x_i(t_i(s)))}] \\
&= \mathbb{E}_{\mathcal{F}^k}[\mathbb{I}_{\mathcal{S}(T)} e^{\gamma}].
\end{aligned}$$

Using A.26 and a change of probability measure from $\mathcal{F}$ to $\mathcal{F}^k$ we have

$$\begin{aligned}
\mathbb{P}_{\mathcal{F}}[\tau_k < \frac{c_1 \log T}{I(f_k, f_k')}, \gamma \leq c_5 \log T] &= \mathbb{E}_{\mathcal{F}}[\mathbb{I}_{\{\tau_k < \frac{c_1 \log T}{I(f_k, f_k')}, \gamma \leq c_5 \log T\}}] \\
&\leq \mathbb{E}_{\mathcal{F}^k}[\mathbb{I}_{\{\tau_k < \frac{c_1 \log T}{I(f_k, f_k')}, \gamma \leq c_5 \log T\}} e^{\gamma}] \\
&\leq T^{c_5} \mathbb{E}_{\mathcal{F}^k}[\mathbb{I}_{\{\tau_k < \frac{c_1 \log T}{I(f_k, f_k')}, \gamma \leq c_5 \log T\}}] \\
&\leq T^{c_5} \mathbb{P}_{\mathcal{F}^k}[\tau_k < \frac{c_1 \log T}{I(f_k, f_k')}] \\
&\leq \frac{T^{c_5} \mathbb{E}_{\mathcal{F}^k}[T - \tau_k]}{T - \frac{c_1 \log T}{I(f_k, f_k')}} \\
&\leq \frac{K T^{c_5 + \alpha}}{T - \frac{c_1 \log T}{I(f_k, f_k')}},
\end{aligned} \tag{A.27}$$

where to arrive at the last inequality we use the $\alpha-$consistency assumption. Form A.22, A.27 and the fact that $|I(f_k, f_*) - I(f_k, f_k')|$ can be arbitrarily small, we conclude that, for $c_5 < 1 - \alpha$

$$\mathbb{P}_{\mathcal{F}}[\tau_k < \frac{c_1 \log T}{I(f_k, f_*)}] \to 0, \ as \ T \to \infty.$$

Equivalently,

$$\mathbb{P}_{\mathcal{F}}[\tau_k \geq \frac{c_1 \log T}{I(f_k, f_*)}] \to 1, \ as \ T \to \infty.$$

Form A.25 and A.27, we conclude that, for $c_5 < 1 - \alpha$, when Assumption 2 is satisfied

$$\mathbb{P}_{\mathcal{F}}[\tau_k < \frac{c_1 \log T}{I(f_k, f_*)}] \leq \frac{c_1 \log T}{I(f_k, f_k')} T^{-a_1 I(f_k, f_k') \frac{(c_5 - c_1)^2}{c_1}} + \frac{K T^{c_5 + \alpha}}{T - \frac{c_1 \log T}{I(f_k, f_k')}}.$$

Thus, there is a $T_0 \in \mathbb{N}$ such that for $T \geq T_0$,

$$\mathbb{P}_{\mathcal{F}}[\tau_k \geq \frac{c_1 \log T}{I(f_k, f_*)}] \geq c_2.$$

for some constant $c_2 > 0$ independent of T and $\mathcal{F}$. We emphasize that the constant c_1 and c_5 are chosen to satisfy $c_1 < c_5 < 1 - \alpha$.

A.6 Proof of Theorem 13

Since $R_\pi(T) \geq \widehat{R}_\pi(T)$ we can establish a lower bound on $\widehat{R}_\pi(T)$. From Lemma 3 we have

$$\widehat{R}_\pi(T) \quad = \quad \sum_{i=1}^{K} \mathbb{E}\tau_i \Delta_i + \sum_{i=1}^{K} \mathbb{E}\tau_i \Gamma_{i,*}^2 - \frac{1}{T} \mathbb{E}[(\sum_{i=1}^{K} \tau_i (\overline{\mu}_i - \mu_*))^2] + \sigma_*^2. \qquad \text{(A.28)}$$

A lower bound on $\mathbb{E}[\tau_i]$ is a straightforward consequence of Lemma 5. By Markov inequality we have

$$\mathbb{E}[\tau_i] \geq \mathbb{P}[\tau_i \geq \frac{c_1 \log T}{I(f_i, f_*)}]\frac{c_1 \log T}{I(f_i, f_*)}. \tag{A.29}$$

So we can write

$$\begin{aligned} \lim_{T\to\infty} \frac{\mathbb{E}[\tau_i]}{\log T} &\geq \lim_{T\to\infty} \mathbb{P}[\tau_i \geq \frac{c_1 \log T}{I(f_i, f_*)}]\frac{c_1}{I(f_i, f_*)} \\ &= \frac{c_1}{I(f_i, f_*)}. \end{aligned} \tag{A.30}$$

Similarly, by Markov inequality we have, when Assumption 2 is satisfied, there is a $T_0 \in \mathbb{N}$ such that for all $T \geq T_0$,

$$\mathbb{E}[\tau_i] \geq \frac{c_1 c_2 \log T}{I(f_i, f_*)}. \tag{A.31}$$

For the third term on the RHS of regret expression A.28, following the similar steps as in A.19, we have

$$\frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K} \tau_i(\overline{\mu}_i - \mu_*))^2] \leq \sigma_{\max}^2 + \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K} \tau_i \Gamma_{i,*})^2] + 2\sigma_{\max} \sum_{i=1}^{K} \sqrt{\mathbb{E}\tau_i \Gamma_{i,*}^2}. \tag{A.32}$$

Define the event $\mathcal{E}$ as follows. $\mathcal{E}$: for all $k \neq *$, $\tau_k \leq T^{\frac{1+\alpha}{2}}$.

For $\sum_{i=1}^{K} \mathbb{E}\tau_i \Gamma_{i,*}^2 - \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K} \tau_i \Gamma_{i,*})^2]$, we have

$$\begin{aligned} \sum_{i=1}^{K} \mathbb{E}\tau_i \Gamma_{i,*}^2 - \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K} \tau_i \Gamma_{i,*})^2] &= \sum_{i=1}^{K} \mathbb{E}[\tau_i \Gamma_{i,*}^2 - \frac{1}{T}(\sum_{i=1}^{K} \tau_i \Gamma_{i,*})^2] && \text{(A.33)} \\ &\geq \sum_{i=1}^{K} \mathbb{E}[\tau_i \Gamma_{i,*}^2 - \frac{1}{T}(\sum_{i=1}^{K} \tau_i \Gamma_{i,*})^2, \mathcal{E}] && \text{(A.34)} \\ &= \sum_{i\neq *} \mathbb{E}[\tau_i(\Gamma_{i,*}^2 - \frac{1}{T}\sum_{j\neq *} \tau_j \Gamma_{i,*}\Gamma_{j,*}), \mathcal{E}] \end{aligned}$$

$$\begin{aligned} &\geq \sum_{i\neq *} \mathbb{E}[\tau_i(\Gamma_{i,*}^2 - \frac{1}{T}(K-1)\Gamma^2 T^{\frac{1+\alpha}{2}}), \mathcal{E}] \\ &= \sum_{i\neq *} \mathbb{E}[\tau_i(\Gamma_{i,*}^2 - (K-1)\Gamma^2 T^{-\frac{1-\alpha}{2}}), \mathcal{E}]. \quad \text{(A.35)} \end{aligned}$$

Notice that A.34 holds because the argument inside the expectation in A.33 is always positive (similar to A.18 by Cauchy-Schwartz inequality).

We also have

$$\begin{aligned} \mathbb{P}[\tau_i \geq \frac{c_1 \log T}{I(f_i, f_*)}, \mathcal{E}] &= \mathbb{P}[\tau_i \geq \frac{c_1 \log T}{I(f_i, f_*)}] - \mathbb{P}[\tau_i < \frac{c_1 \log T}{I(f_i, f_*)}, \overline{\mathcal{E}}] \\ &\geq \mathbb{P}[\tau_i \geq \frac{c_1 \log T}{I(f_i, f_*)}] - \mathbb{P}[\overline{\mathcal{E}}] \\ &\geq \mathbb{P}[\tau_i \geq \frac{c_1 \log T}{I(f_i, f_*)}] - \sum_{i\neq *} \mathbb{P}[\tau_i > T^{\frac{1+\alpha}{2}}] \\ &\geq \mathbb{P}[\tau_i \geq \frac{c_1 \log T}{I(f_i, f_*)}] - \sum_{i\neq *} \frac{\mathbb{E}[\tau_i]}{T^{\frac{1+\alpha}{2}}} \\ &\geq \mathbb{P}[\tau_i \geq \frac{c_1 \log T}{I(f_i, f_*)}] - (K-1)T^{-\frac{1-\alpha}{2}}. \quad \text{(A.36)} \end{aligned}$$

Thus, by Markov inequality

$$\begin{aligned} \mathbb{E}[\tau_i, \mathcal{E}] &\geq \mathbb{P}[\tau_i \geq \frac{c_1 \log T}{I(f_i, f_*)}, \mathcal{E}]\frac{c_1 \log T}{I(f_i, f_*)} \\ &\geq \mathbb{P}[\tau_i \geq \frac{c_1 \log T}{I(f_i, f_*)}]\frac{c_1 \log T}{I(f_i, f_*)} - (K-1)T^{-\frac{1-\alpha}{2}}\frac{c_1 \log T}{I(f_i, f_*)} \end{aligned}$$

As a result of Lemma 5, we have

$$\begin{aligned} \lim_{T\to\infty} \frac{\mathbb{E}[\tau_i, \mathcal{E}]}{\log T} &\geq \lim_{T\to\infty} \mathbb{P}[\tau_i \geq \frac{c_1 \log T}{I(f_i, f_*)}]\frac{c_1}{I(f_i, f_*)} \quad \text{(A.37)} \\ &\quad - \lim_{T\to\infty} \frac{c_1(K-1)T^{-\frac{1-\alpha}{2}}}{I(f_i, f_*)} \\ &= \frac{c_1}{I(f_i, f_*)}. \end{aligned}$$

Similarly, we have, when Assumption 2 is satisfied, there is a $T_0 \in \mathbb{N}$ such that for all $T \geq T_0$,

$$\mathbb{E}[\tau_k, \mathcal{E}] \geq \frac{c_1 c_2 \log T}{I(f_i, f_*)} - (K-1)T^{-\frac{1-\alpha}{2}}\frac{c_1 \log T}{I(f_i, f_*)}. \quad \text{(A.38)}$$

Substituting A.32 and A.35 in regret expression we have

$$\begin{aligned}\widehat{R}_\pi(T) &\geq \sum_{i=1}^{K} \mathbb{E}\tau_i\Delta_i + \sum_{i\neq *} \mathbb{E}[\tau_i, \mathcal{E}](\Gamma_{i,*}^2 - (K-1)\Gamma^2 T^{-\frac{1-\alpha}{2}}) \\ &\quad - 2\sigma_{\max}\sum_{i=1}^{K}\sqrt{\mathbb{E}\tau_i\Gamma_{i,*}^2} - \sigma_{\max}^2 \\ &= \sum_{i=1}^{K} \mathbb{E}\tau_i(\Delta_i - \frac{2\sigma_{\max}|\Gamma_{i,*}|}{\sqrt{\mathbb{E}\tau_i}} - \frac{\sigma_{\max}^2}{\mathbb{E}\tau_i}) \\ &\quad + \sum_{i\neq *} \mathbb{E}[\tau_i, \mathcal{E}](\Gamma_{i,*}^2 - (K-1)\Gamma^2 T^{-\frac{1-\alpha}{2}}).\end{aligned}$$

Substituting the lower bounds on $\mathbb{E}[\tau_i]$ and $\mathbb{E}[\tau_i, \mathcal{E}]$ in the above bound we arrive at

$$\liminf_{T\to\infty} \frac{\widehat{R}_\pi(T)}{\log T} \geq \sum_{i\neq *} \frac{c_1}{I(f_i, f_*)}(\Delta_i + \Gamma_{i,*}^2). \tag{A.39}$$

Also, when Assumption 2 is satisfied, for $T \geq T_0$,

$$\begin{aligned}\widehat{R}_\pi(T) &\geq \sum_{i\neq *} \frac{c_1 c_2 \log T}{I(f_i, f_*)}(\Delta_i - \frac{2\sigma_{\max}|\Gamma_{i,*}|}{\sqrt{\frac{c_1 c_2 \log T}{I(f_i,f_*)}}} - \frac{\sigma_{\max}^2}{\frac{c_1 c_2 \log T}{I(f_i,f_*)}}) \\ &\quad + \sum_{i\neq *}(\frac{c_1 c_2 \log T}{I(f_i, f_*)} - (K-1)T^{-\frac{1-\alpha}{2}}\frac{c_1 \log T}{I(f_i, f_*)}) \\ &\quad (\Gamma_{i,*}^2 - (K-1)\Gamma^2 T^{-\frac{1-\alpha}{2}}).\end{aligned}$$

We can rewrite the above lower bound such that for all $T \geq T_1$,

$$\widehat{R}_\pi(T) \geq \sum_{i\neq *} \frac{c_1 c_2 \log T}{I(f_i, f_*)}(\Delta_i + \Gamma_{i,*}^2 - \epsilon_{T_1}), \tag{A.40}$$

where ϵ_{T_1} can be arbitrary small when T_1 is large enough. However, precise characterization of ϵ_{T_1} is tedious and depends on all of the diminishing terms above. ■

A.7 Proof of Lemma 6

Let $i \neq *$ be a suboptimal arm and $b_i = \lceil \frac{4b^2 \log T}{\min\{\Delta_i^2, 4(2+\rho)^2\}} \rceil$.

$$\begin{aligned}
\tau_i(T) &= \sum_{t=1}^{T} \mathbb{I}[\pi(t) = i] \\
&= b_i + \sum_{t=b_i+1}^{T} \mathbb{I}[\pi(t) = i, \tau_i(t) \geq b_i] \\
&\leq b_i + \sum_{t=b_i+1}^{T} \mathbb{I}[\eta_i(t) - \eta_*(t) \leq 0, \tau_i(t) \geq b_i]. \qquad \text{(A.41)}
\end{aligned}$$

We can write

$$\begin{aligned}
\eta_i(t) - \eta_*(t) &= \overline{\xi}_i(t) - b\sqrt{\frac{\log t}{\tau_i(t)}} - \eta_*(t) \\
&= (\overline{\xi}_i(t) + b\sqrt{\frac{\log t}{\tau_i(t)}} - \xi_i) - (\eta_*(t) - \xi_*) \\
&+ (\xi_i - \xi_* - 2b\sqrt{\frac{\log t}{\tau_i(t)}}). \qquad \text{(A.42)}
\end{aligned}$$

For $\tau_i(t) \geq b_i$, the last term in A.42 is positive, thus continuing from A.41

$$\begin{aligned}
\tau_i(T) &\leq b_i + \sum_{t=b_i+1}^{T} \mathbb{I}[\overline{\xi}_i(t) + b\sqrt{\frac{\log t}{\tau_i(t)}} - \xi_i \leq 0, \tau_i(t) \geq b_i] \\
&+ \sum_{t=b_i+1}^{T} \mathbb{I}[\eta_*(t) - \xi_* \geq 0].
\end{aligned}$$

Applying Lemma 2

$$\begin{aligned}
\mathbb{E}[\tau_i(T)] &\leq b_i + 2\sum_{t=b_i+1}^{T} t\exp(-\frac{ab^2 \log t}{(2+\rho)^2}) + 2\sum_{t=b_i+1}^{T} t\exp(-\frac{ab^2 \log t}{(1+\rho)^2}) \\
&\leq \frac{4b^2 \log T}{\min\{\Delta_i^2, 4(2+\rho)^2\}} + 1 + 4\int_{b_i}^{\infty} t^{-2} dt \\
&= \frac{4b^2 \log T}{\min\{\Delta_i^2, 4(2+\rho)^2\}} + 1 + 4b_i^{-1} \\
&\leq \frac{4b^2 \log T}{\min\{\Delta_i^2, 4(2+\rho)^2\}} + 5. \quad \blacksquare \qquad \text{(A.43)}
\end{aligned}$$

A.8 Proof of Theorem 14

Considering the regret expression in Lemma 3

$$\begin{aligned}\widehat{R}_{MV-UCB}(T) &= \sum_{i=1}^{K}\mathbb{E}\tau_i\Delta_i + \sum_{i=1}^{K}\mathbb{E}\tau_i\Gamma_{i,*}^2 - \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K}\tau_i(\overline{\mu}_i-\mu_*))^2] + \sigma_*^2 \\ &\leq \sum_{i\neq *}\mathbb{E}\tau_i(\Delta_i+\Gamma_{i,*}^2) + \sigma_*^2 \\ &\leq \sum_{i\neq *}(\frac{4b^2\log T}{\Delta_i^2}+5)(\Delta_i+\Gamma_{i,*}^2) + \sigma_*^2.\end{aligned}$$

From Theorem 12, we have

$$R_{MV-UCB}(T) \leq \sum_{i\neq *}(\frac{4b^2\log T}{\Delta_i^2}+5)(\Delta_i+\Gamma_{i,*}^2) + \sigma_*^2 + \min\{\sigma_{\max}^2(\sum_{i\neq *}\frac{\Gamma_{i,*}^2}{\Delta_i}+1), \frac{K}{a}\log T\}. \quad \blacksquare$$

A.9 Proof of Theorem 16

The following lemma is used in the proof of the theorem.

Lemma 10 ([11]) *Let ν_0, ν_1 be two probability measures supported on some set $\mathcal{X}$, with ν_1 absolutely continuous with respect to ν_0. Then, for any measurable function $\phi : \mathcal{X} \to \{0,1\}$,*

$$\mathbb{P}_{\nu_1}[\phi(X)=0] + \mathbb{P}_{\nu_0}[\phi(X)=1] \geq \frac{1}{2}\exp(-I(\nu_0,\nu_1)). \tag{A.44}$$

To prove Theorem 16, two different sets of distributions are assigned to a two-armed bandit. Then it is shown that under at least one of these two sets of distributions 5.22 holds. Consider a two-armed bandit. Let $f_1 \sim \mathcal{N}(\mu_1,\sigma_1^2)$, a normal distribution with mean $\mu_1 = \frac{3}{4}$ and variance $\sigma_1^2 = \frac{3}{16} - 4\Delta^2$. Also, let $f_2 \sim \mathcal{B}(p)$, a Bernolli distribution with $p = 1/4 + 2\Delta$ and $f_2' \sim \mathcal{B}(q)$ with $q =$

$1/4 - 2\Delta$. Denote $\mathcal{F} = (f_1, f_2)$ and $\mathcal{F}' = (f_1, f_2')$. For the simplicity of presentation let us assume $\rho = 0$. Note that for the difference between the variance of above distributions we have $\sigma_2^2 - \sigma_1^2 = \Delta$ and $\sigma_1^2 - \sigma_2'^2 = \Delta$. Since $\widehat{R}_\pi(T) \leq R_\pi(T)$ we can establish a lower bound on $\widehat{R}_\pi(T)$ that is also a lower bound on $R_\pi(T)$. From Lemma 3

$$\widehat{R}_\pi(T) = \sum_{i=1}^{K} \mathbb{E}\tau_i\Delta_i + \sum_{i=1}^{K} \mathbb{E}\tau_i\Gamma_{i,*}^2 - \frac{1}{T}\mathbb{E}[(\sum_{i=1}^{K} \tau_i(\overline{\mu}_i - \mu_*))^2] + \sigma_*^2$$

Following the similar lines as the proof of A.21 we show

$$\widehat{R}_\pi(T) \geq \sum_{i=1}^{2} \mathbb{E}\tau_i\Delta_i - \frac{2}{a}(\log T + 2).$$

Using Lemma 10 through a coupling argument we establish a lower bound on the regret under one of the two systems.

Let us use the notations $R_\pi(T;\mathcal{F})$ and $R_\pi(T;\mathcal{F}')$ to distinguish between the regrets under distribution assignments $\mathcal{F}$ and $\mathcal{F}'$, respectively. Also, let $f^{(t)}$ and $f'^{(t)}$ denote the distribution of the reward process up to time t under $\mathcal{F}$ and $\mathcal{F}'$, respectively. Spcifically,

$$f^{(t)}(x(1), x(2), ..., x(t)) = \Pi_{\{s:\pi(s)=1\}} f_1(x(s))\Pi_{\{s:\pi(s)=2\}} f_2(x(s))$$

and $f'^{(t)}$ is defined similarly under $\mathcal{F}'$.

$$\begin{aligned}
&\max(\widehat{R}_\pi(T;\mathcal{F}), \widehat{R}_\pi(T;\mathcal{F}')) \\
&\quad \geq \frac{1}{2}(\widehat{R}_\pi(T;\mathcal{F})) + \widehat{R}_\pi(T;\mathcal{F}')) \\
&\quad \geq \frac{\Delta}{2}\sum_{t=1}^{T}(\mathbb{P}_{\mathcal{F}}[\pi(t) = 2] + \mathbb{P}_{\mathcal{F}'}[\pi(t) = 1]) - \frac{2}{a}(\log T + 2)
\end{aligned}$$

$$\geq \frac{\Delta}{2}\sum_{t=1}^{T}(\mathbb{P}_{\mathcal{F}}[\pi(t)=2]+\mathbb{P}_{\mathcal{F}'}[\pi(t)=1])-\frac{2}{a}(\log T+2) \tag{A.45}$$

$$\geq \frac{\Delta}{4}\sum_{t=1}^{T}\exp(-I(f^{(t)},f'^{(t)}))-\frac{2}{a}(\log T+2). \tag{A.46}$$

The KL-divergence between $f^{(t)}$ and $f'^{(t)}$ equals to

$$\begin{aligned} I(f^{(t)},f'^{(t)}) &= \mathbb{E}_{\mathcal{F}}[\log\frac{\Pi_{\{s:\pi(s)=2\}}f_2(X_{\pi(s)}(s))}{\Pi_{\{s:\pi(s)=2\}}f_2'(X_{\pi(s)}(s))}] \\ &= \mathbb{E}_{\mathcal{F}}[\sum_{s=1}^{\tau_2(t)}(p\log\frac{p}{q}+(1-p)\log\frac{1-p}{1-q})] \\ &= \mathbb{E}_{\mathcal{F}}\tau_2(t)(p\log\frac{p}{q}+(1-p)\log\frac{1-p}{1-q}) \\ &\leq \mathbb{E}_{\mathcal{F}}\tau_2(t)d_0\Delta^2. \end{aligned} \tag{A.47}$$

for some constant d_0. Substituting A.47 in A.46

$$\begin{aligned} &\max(\widehat{R}_\pi(T;\mathcal{F}),\widehat{R}_\pi(T;\mathcal{F}')) \\ &\quad\geq \frac{\Delta}{4}T\exp(-\mathbb{E}_{\mathcal{F}}\tau_2(T)d_0\Delta^2)-\frac{2}{a}(\log T+2). \end{aligned} \tag{A.48}$$

Following the similar lines as the proof of A.17, we show that

$$\begin{aligned} \widehat{R}_\pi(T) \geq &\sum_{i=1}^{2}\mathbb{E}\tau_i\Delta_i+\sum_{i=1}^{2}\mathbb{E}\tau_i\Gamma_{i,*}^2-\frac{1}{T}\mathbb{E}[(\sum_{i=1}^{2}\tau_i\Gamma_{i,*})^2] \\ &-\sigma_{\max}^2-2\sigma_{\max}\sqrt{\frac{1}{T}\mathbb{E}[(\sum_{i=1}^{2}\tau_i\Gamma_{i,*})^2]}. \end{aligned}$$

We can write

$$\begin{aligned} \widehat{R}_\pi(T;\mathcal{F}) &\geq \mathbb{E}_{\mathcal{F}}\tau_2\Delta+\mathbb{E}_{\mathcal{F}}\tau_2\Gamma^2-\frac{1}{T}\mathbb{E}_{\mathcal{F}}[\tau_2^2]\Gamma^2-\sigma_{\max}^2-2\sigma_{\max}\sqrt{\frac{1}{T}\mathbb{E}_{\mathcal{F}}[\tau_2^2]\Gamma^2} \\ &= \mathbb{E}_{\mathcal{F}}\tau_2\Delta+\frac{1}{T}\mathbb{E}_{\mathcal{F}}[\tau_1\tau_2]\Gamma^2-\sigma_{\max}^2-2\sigma_{\max}\sqrt{\mathbb{E}_{\mathcal{F}}[\tau_2]\Gamma^2}. \end{aligned} \tag{A.49}$$

For the first term on the right hand side of A.49, we have, for a constant $0 < d_3 < 1$

$$\begin{aligned}\frac{1}{T}\mathbb{E}_{\mathcal{F}}[\tau_1(T)\tau_2(T)] &\geq \frac{T-d_3T}{T}\mathbb{E}_{\mathcal{F}}[\tau_2, \tau_2 \leq d_3T] \\ &\geq \frac{1}{2}\frac{T-d_3T}{T}\mathbb{E}_{\mathcal{F}}[\tau_2] \qquad \text{(A.50)} \\ &= d_4\mathbb{E}_{\mathcal{F}}[\tau_2], \qquad \text{(A.51)}\end{aligned}$$

where $d_4 = \frac{1-d_3}{2}$. To arrive at A.50, notice that we have

$$\mathbb{E}_{\mathcal{F}}[\tau_2, \tau_2 > d_3T] \leq \mathbb{E}_{\mathcal{F}}[\tau_2, \tau_2 \leq d_3T], \qquad \text{(A.52)}$$

otherwise $\mathbb{E}_{\mathcal{F}}[\tau_2]$ is linear with time and we arrive at the theorem. We show that $\mathbb{E}_{\mathcal{F}}[\tau_2, \tau_2 > d_3T] \leq \mathbb{E}_{\mathcal{F}}[\tau_2, \tau_2 \leq d_3T]$ translates to $\mathbb{E}_{\mathcal{F}}[\tau_2, \tau_2 \leq d_3T] \geq \frac{1}{2}\mathbb{E}[\tau_2]$.

$$\begin{aligned}&\mathbb{E}_{\mathcal{F}}[\tau_2] - \mathbb{E}[\tau_2, \tau_2 > d_3T] \\ &= \mathbb{E}_{\mathcal{F}}[\tau_2, \tau_2 \leq d_3T] + \mathbb{E}_{\mathcal{F}}[\tau_2, \tau_2 > d_3T] - \mathbb{E}_{\mathcal{F}}[\tau_2, \tau_2 > d_3T] \\ &\geq \mathbb{E}_{\mathcal{F}}[\tau_2, \tau_2 \leq d_3T] + \mathbb{E}_{\mathcal{F}}[\tau_2, \tau_2 > d_3T] - \frac{1}{2}\mathbb{E}_{\mathcal{F}}[\tau_2, \tau_2 > d_3T] - \frac{1}{2}\mathbb{E}_{\mathcal{F}}[\tau_2, \tau_2 \leq d_3T] \\ &= \frac{1}{2}(\mathbb{E}_{\mathcal{F}}[\tau_2, \tau_2 \leq d_3T] + \mathbb{E}_{\mathcal{F}}[\tau_2, \tau_2 > d_3T]) \\ &= \frac{1}{2}\mathbb{E}_{\mathcal{F}}[\tau_2].\end{aligned}$$

By A.49 and A.51, we can write

$$\begin{aligned}\max(\widehat{R}_\pi(T;\mathcal{F}), \widehat{R}_\pi(T;\mathcal{F}')) &\geq \widehat{R}_\pi(T;\mathcal{F}) \\ &\geq \mathbb{E}_{\mathcal{F}}\tau_2\Delta + d_4\mathbb{E}_{\mathcal{F}}\tau_2\Gamma^2 - \sigma_{\max}^2 - 2\sigma_{\max}\sqrt{\mathbb{E}_{\mathcal{F}}[\tau_2]\Gamma^2} \qquad \text{(A.53)}\end{aligned}$$

For brevity of notation, let $x \triangleq \mathbb{E}_{\mathcal{F}}\tau_2$. From A.48 and A.53 (by taking average over two lower bounds) we have, for $T \geq T_0$ for some large enough $T_0 \in \mathbb{N}$

$$\max(R_\pi(T;\mathcal{F}), R_\pi(T;\mathcal{F}')) \geq \frac{1}{2}\{\frac{T\Delta}{4}\exp(-\mathbb{E}_{\mathcal{F}}\tau_2 d_0\Delta^2) - \frac{2}{a}(\log T + 2) + \mathbb{E}_{\mathcal{F}}\tau_2\Delta$$

$$
\begin{aligned}
&+ d_4 \mathbb{E}_{\mathcal{F}} \tau_2 \Gamma^2 - \sigma_{\max}^2 - 2\sigma_{\max} \sqrt{\mathbb{E}_{\mathcal{F}} \tau_2 \Gamma^2}\} \\
\geq\ & \min_{x \geq 0} \frac{1}{2} \{ \frac{T\Delta}{4} \exp(-x d_0 \Delta^2) + d_4 x \Gamma^2 \} \\
&+ \min_{x \geq 0} \frac{1}{2} \{ x\Delta - 2\sigma_{\max} \sqrt{x\Gamma^2} \} \\
&- \frac{1}{a}(\log T + 2) - \frac{1}{2}\sigma_{\max}^2 \\
=\ & \frac{d_2 \Gamma^2}{2 d_0 \Delta^2} (\log \frac{T d_0 \Delta^3}{4 d_4 \Gamma^2} + 1) \\
&- \frac{\sigma_{\max}^2 \Gamma^2}{2\Delta} - \frac{1}{a}(\log T + 2) - \frac{1}{2}\sigma_{\max}^2 \qquad \text{(A.54)} \\
\geq\ & \frac{c_4 \log T}{\Delta^2}.
\end{aligned}
$$

Substituting Δ with $d_6 T^{-1/3}$ in A.54 for a constant d_6 that satisfies $\frac{d_0 d_6^3}{4 d_4 \Gamma^2} > 1$, we have, for some constant $c_3 > 0$,

$$
R_\pi(T) \geq c_3 T^{2/3}. \qquad \text{(A.55)}
$$

In this proof, for the purpose of presentation, it is assumed $\rho = 0$. For $\rho \neq 0$ the same proof holds with modified assignment of distributions. The assignment of distributions are as follows. For $\rho \neq \frac{1}{2}$, let $p = \frac{1}{4} + 2\delta$, $q = \frac{1}{4} - 2\delta$, $\mu_1 = \frac{3}{4}$ and $\sigma_1^2 = \frac{3}{16} - 4\delta^2 + \frac{\rho}{2}$. For $\rho = \frac{1}{2}$, let $p = \frac{1}{3} + 3\delta$, $q = \frac{1}{3} - 3\delta$, $\mu_1 = \frac{5}{6}$ and $\sigma_1^2 = \frac{17}{36} - 9\delta^2$. ■

A.10 Proof of Theorem 8

To prove Theorem 8 we consider a K-armed bandit where the rewards assigned to the arms have Bernoulli distribution. Let $\mathcal{B}(p)$ denote a Bernoulli distribution with mean value p. Consider three different settings. First, the reward on each arm has a uniform Bernoulli distribution, $X_i(t) \sim \mathcal{B}(\frac{1}{2})$. We denote by $\mathbb{E}_0$ and $\mathbb{P}_0$ the expectation operator and probability measure under this

setting, respectively. Second, except arm i the reward on each arm has a uniform Bernoulli distribution. The reward on arm i is the following stochastic process: $X_i(t) \sim \mathcal{B}(\frac{1}{2} + \epsilon_t)$ with $\epsilon_t > 0$. We denote by $\mathbb{E}_i$ and $\mathbb{P}_i$ the expectation operator and probability measure under this setting, respectively. The third setting is similar to the second setting except arm i is chosen randomly from K arms with equal probability. We denote by $\mathbb{E}_u$ and $\mathbb{P}_u$ the expectation operator and probability measure under this setting, respectively.

Following the same approach as in [12] we can upper bound the difference between the probabilities of selecting arm i at time t under first two settings. Let $Z(t) = X_{\pi(t)}(t)$ the reward obtained at time t by an arm selection policy π. Let $Z^{(t)} = [Z(1), ..., Z(t)]$ be the sequence of rewards received up to time t.

$$\begin{aligned}
& \mathbb{P}_i(\pi(t) = i) - \mathbb{P}_0(\pi(t) = i) \\
= & \sum_{Z^{(t-1)} \in \{0,1\}^{t-1}} \mathbb{I}(\pi(t) = i)(\mathbb{P}_i[Z^{(t-1)}] - \mathbb{P}_0[Z^{(t-1)}]) \\
\leq & \sum_{\substack{Z^{(t-1)} \in \{0,1\}^{t-1}, \\ \mathbb{P}_i[Z^{(t-1)}] \geq \mathbb{P}_0[Z^{(t-1)}]}} \mathbb{I}(\pi(t) = i)(\mathbb{P}_i[Z^{(t-1)}] - \mathbb{P}_0[Z^{(t-1)}]) \\
= & \frac{1}{2}||\mathbb{P}_i - \mathbb{P}_0||. \qquad \text{(A.56)}
\end{aligned}$$

It is known that [39]

$$\frac{1}{2}||\mathbb{P}_i - \mathbb{P}_0||^2 \leq (2 \ln 2)\mathcal{D}(\mathbb{P}_0||\mathbb{P}_i), \qquad \text{(A.57)}$$

where $\mathcal{D}(.||.)$ denotes the Kullback-Liebler divergence between two distributions. The value of $\mathcal{D}(\mathbb{P}_0||\mathbb{P}_i)$ can be calculated as

$$\begin{aligned}
\mathcal{D}(\mathbb{P}_0||\mathbb{P}_i) &= \sum_{t=1}^{T} \mathcal{D}(\mathbb{P}_0(X_{\pi(t)}|X^{t-1})||\mathbb{P}_i(X_{\pi(t)}|X^{t-1})) \\
&= \sum_{t=1}^{T} \mathbb{P}_0(\pi(t) = i)\mathcal{D}(\mathcal{B}(\frac{1}{2})||\mathcal{B}(\frac{1}{2} + \epsilon_t))
\end{aligned}$$

$$= \sum_{t=1}^{T} \mathbb{P}_0(\pi(t) = i)(-\frac{1}{2}\log(1-4\epsilon_t^2)). \quad (A.58)$$

From A.56, A.57 and A.58, we have

$$\mathbb{P}_i(\pi(t) = i) \leq \mathbb{P}_0(\pi(t) = i) + \frac{1}{2}\sqrt{-\log 2 \sum_{t=1}^{T} \mathbb{P}_0(\pi(t) = i)(\log(1-4\epsilon_t^2))}. \quad (A.59)$$

Since

$$\mathbb{E}_i[\sum_{t=1}^{T} X_{\pi(t)}] = \frac{T}{2} + \sum_{t=1}^{T} \mathbb{P}_i(\pi(t) = i)\epsilon_t, \quad (A.60)$$

taking average over $i = 1, ..., K$, we have

$$\begin{aligned}
\mathbb{E}_u[\sum_{t=1}^{T} X_{\pi(t)}] &= \frac{1}{K}\sum_{i=1}^{K} \mathbb{E}_i[\sum_{t=1}^{T} X_{\pi(t)}] \\
&= \frac{T}{2} + \frac{1}{K}\sum_{i=1}^{K}\sum_{t=1}^{T} \mathbb{P}_i(\pi(t) = i)\epsilon_t \\
&\leq \frac{T}{2} + \frac{1}{K}\sum_{i=1}^{K}\sum_{t=1}^{T} \mathbb{P}_0(\pi(t) = i)\epsilon_t \\
&+ \frac{1}{2K}\sum_{i=1}^{K}\sum_{t=1}^{T} \epsilon_t \sqrt{-\log 2 \sum_{t=1}^{T} \mathbb{P}_0(\pi(t) = i)\log(1-4\epsilon_t^2)} \quad (A.61) \\
&\leq \frac{T}{2} + \frac{1}{K}\sum_{t=1}^{T} \epsilon_t \\
&+ \frac{1}{2\sqrt{K}}\sum_{t=1}^{T} \epsilon_t \sqrt{-\log 2 \sum_{i=1}^{K}\sum_{t=1}^{T} \mathbb{P}_0(\pi(t) = i)\log(1-4\epsilon_t^2)} \quad (A.62) \\
&\leq \frac{T}{2} + \frac{1}{K}\sum_{t=1}^{T} \epsilon_t \\
&+ \frac{1}{2\sqrt{K}}\sum_{t=1}^{T} \epsilon_t \sqrt{-\log 2 \sum_{t=1}^{T} (\log(1-4\epsilon_t^2))} \quad (A.63) \\
&\leq \frac{T}{2} + \frac{1}{K}\sum_{t=1}^{T} \epsilon_t \\
&+ \frac{1}{2\sqrt{K}}\sum_{t=1}^{T} \epsilon_t \sqrt{\log 2 \sum_{t=1}^{T} 4\epsilon_t^2}. \quad (A.64)
\end{aligned}$$

Notice that $\sum_{i=1}^{K} \mathbb{P}_0(\pi(t) = i) = 1$. The inequality A.61 is obtained from A.59. The inequality A.62 is a result of inequality $\frac{1}{\sqrt{K}} \sum_{i=1}^{K} \sqrt{\alpha_i} \leq \sqrt{\sum_{i=1}^{K} \alpha_i}$. Since $\log(1-\alpha) \geq -\alpha$ for positive α we arrive at A.64. Let $\overline{R}_{\pi}^{w}(T)$ denote the average weak regret when the optimal arm is chosen randomly according to the third setting. Let

$$\epsilon_t = \frac{b\sqrt{K}}{t^{\frac{1}{2}+\delta}}, \tag{A.65}$$

for some small positive b and δ. From A.64 we have

$$\begin{aligned}
\overline{R}_{\pi}^{w}(T) &\geq \sum_{t=1}^{T}(1-\frac{1}{K})\epsilon_t - \frac{1}{2\sqrt{K}}\sum_{t=1}^{T}\epsilon_t\sqrt{\log 2\sum_{t=1}^{T}4\epsilon_t^2} \\
&\geq (1-\frac{1}{K})\epsilon_t \\
&- \frac{1}{2\sqrt{K}}\sum_{t=1}^{T}\epsilon_t\sqrt{4\log 2Kb^2(1+\frac{1}{2\delta}(1-T^{-2\delta}))} \\
&\geq (1-\frac{1}{K})\frac{b\sqrt{K}}{\frac{1}{2}-\delta}(T^{\frac{1}{2}-\delta}-1) \\
&- \frac{b}{2}(1+\frac{1}{\frac{1}{2}-\delta}(T^{\frac{1}{2}-\delta}-1))\sqrt{4\log 2Kb^2(1+\frac{1}{2\delta}(1-T^{-2\delta}))} \\
&= \Theta(b\sqrt{K}T^{\frac{1}{2}-\delta} - \frac{b^2\sqrt{K}}{\sqrt{\delta}}T^{\frac{1}{2}-2\delta}).
\end{aligned} \tag{A.66}$$

Equation A.66 shows that for at least one of the cases in the second setting

$$R_{\pi}^{w}(T) = \Theta(b\sqrt{K}T^{\frac{1}{2}-\delta} - \frac{b^2\sqrt{K}}{\sqrt{\delta}}T^{\frac{1}{2}-2\delta}). \tag{A.67}$$

Since δ is any arbitrary positive number, we can let δ to converge zero. With the choice of a small value for b the lower bound converges to $\Theta(\sqrt{KT})$. ■

A.11 Proof of Theorem 23

It is shown in [12] that for a deterministic sequence of rewards $Z^{(T)}$,

$$G_{EXP3}(Z^{(T)}) \geq (1-(e-1)\gamma)G_{\max}(Z^{(T)}) - \frac{K\log K}{\gamma}, \tag{A.68}$$

where $G_{EXP3}(Z^{(T)})$ is the total reward obtained by EXP3 algorithm and $G_{\max}(Z^{(T)})$ is the total reward obtained by a genie who is restricted to play a single arm over the entire time horizon. Taking expectation on the distribution of arms from both sides

$$\mathbb{E}[G_{EXP3}(X^{(T)})] \geq (1-(e-1)\gamma)\mathbb{E}[G_{\max}(X^{(T)})] - \frac{K\log K}{\gamma}.$$

Since $\mathbb{E}[G_{\max}(X^{(T)})] \geq \max_i \sum_{t=1}^{T} \mu_i(t)$, we have

$$\mathbb{E}[G_{EXP3}(X^{(T)})] \geq (1-(e-1)\gamma)\max_i \sum_{t=1}^{T} \mu_i(t) - \frac{K\log K}{\gamma}.$$

Thus,

$$R^{w}_{EXP3}(T) \leq (e-1)\gamma \max_i \sum_{i=1}^{T} \mu_i(t) + \frac{K\log K}{\gamma}.$$

Since $\sum_{i=1}^{T} \mu_i(t) \leq T$ for all i, with proper choice of $\gamma = \sqrt{\frac{K\log K}{(e-1)T}}$, we arrive at theorem

$$R^{w}_{EXP3}(T) \leq 2\sqrt{(e-1)TK\log K}. \quad \blacksquare$$

A.12 Proof of Theorem 10

From [12] we know, for a deterministic sequence of rewards $Z^{(T)}$ where the genie switches no more than $H(T)-1$ times across the arms,

$$G_{\max}(Z^{(T)}) - G_{EXP3.S}(Z^{(T)}) \leq \frac{K(H(T)\log(K/\alpha) + e\alpha T)}{\gamma} + (e-1)\gamma T,$$

where $G_{\max}(Z^{(T)})$ and $G_{EXP3.S}(Z^{(T)})$ are the rewards obtained by genie and EXP3.S algorithm, respectively. Let $i^{(T)} = \{i(1), i(2), ..., i(T)\} \in \{1, ..., M\}^T$ denote a sequence of played arms. Let $\mathcal{S}_{H(T)}$ denote all such sequences consisting of up

to $H(T)$ segments where a single arm is played on each segment. Taking expectation from A.69 and noticing $\mathbb{E}[G_{\max}(Z^{(T)})] \geq \max_{i^{(T)} \in \mathcal{S}_{H(T)}} \sum_{t=1}^{T} \mu_{i(t)}(t)$, we have

$$\max_{i^{(T)} \in \mathcal{S}_{H(T)}} \sum_{t=1}^{T} \mu_{i(t)}(t) - \mathbb{E}[G_{EXP3.S}] \leq \frac{K(H(T)\log(K/\alpha) + e\alpha T)}{\gamma} + (e-1)\gamma T.$$

With proper choice of $\gamma = \sqrt{\frac{K(H(T)\log(KT)+e)}{(e-1)T}}$ and $\alpha = \frac{1}{T}$, we have

$$R_{EXP3.S} \leq 2\sqrt{e-1}\sqrt{KT(H(T)\log(KT)+e)}. \quad \blacksquare$$

A.13 Proof of Theorem 11

If $H(T)$ is known, simply let $\widetilde{H}(T) \triangleq H(T)$. Otherwise, we assume an $\widetilde{H}(T) = \Theta(H(T))$ is given. Let $R_{EXP3S}^{(r)}$ denote the regret on r'th epoch. Thus, $R_{EXP3S}(T) = \sum_{r=0}^{M} R_{EXP3S}^{(r)}$ where M is the epoch containing T. Notice that all the time instances in the last epoch are not necessarily played. Let $L_r = \sum_{i=0}^{r} l_r$. With parameter choice of $\alpha = \frac{1}{l_r}$ and $\gamma = \sqrt{\frac{K(\widetilde{H}(L_r)\log(Kl_r)+e)}{(e-1)l_r}}$ at the beginning of each epoch, from Theorem 10, for $r = 0, 1, ..., M$, and some constant $A > 0$,

$$R_{EXP3S}^{(r)} \leq A\sqrt{K2^r\widetilde{H}(L_r)\log(2^rK)}. \tag{A.69}$$

Thus,

$$\begin{aligned}
R_{EXP3.S}(T) &\leq A\sum_{r=0}^{M}\sqrt{K2^r\widetilde{H}(L_r)\log(2^rK)} \\
&\leq A\sum_{r=0}^{M-1}2^{r/2}\sqrt{K\widetilde{H}(T)\log(TK)} \\
&+ A\sqrt{KT\widetilde{H}(2T)\log(TK)} \qquad \text{(A.70)} \\
&\leq A\frac{2^{M/2}-1}{\sqrt{2}-1}\sqrt{K\widetilde{H}(T)\log(TK)} \\
&+ A\sqrt{2KT\widetilde{H}(T)\log(TK)} \qquad \text{(A.71)}
\end{aligned}$$

$$\leq A\frac{\sqrt{T}-1}{\sqrt{2}-1}\sqrt{K\widetilde{H}(T)\log(TK)}$$

$$+ A\sqrt{2KT\widetilde{H}(T)\log(TK)} \quad \text{(A.72)}$$

$$= O(\sqrt{KTH(T)\log(KT)}). \quad \text{(A.73)}$$

Inequality A.70 holds since $l_r \leq T$ for $r = 0, ..., M$, $L_r \leq T$ for $r = 0, ..., M-1$, and $L_M \leq 2T$. The inequality A.71 holds as a result of the concavity of $\widetilde{H}(T)$. Since $2^M \leq T$, the inequality A.72 holds. ■

APPENDIX B

PROOFS OF LEMMAS AND THEOREMS FROM PART II

B.1 Proof of Theorem 25

For simplicity of notation we use l_k^n, L_k^n and I_k^n instead of $l_k^n(\theta_1, \theta_0)$, $L_k^n(\theta_1, \theta_0)$ and $I_k^n(\theta_1, \theta_0)$, respectively. For λ_n, we have

$$\lambda_n = \frac{1}{\mathbb{P}[\nu > n]} \sum_{k=1}^{n} p_k e^{L_k^n}.$$

Thus,

$$\log \lambda_n \geq L_k^n + \log \frac{p_k}{\mathbb{P}[\nu > n]}. \tag{B.1}$$

For the Shiryaev detection policy

$$\log \lambda_{\tau_S - 1} \leq \log \frac{1-\alpha}{\alpha}. \tag{B.2}$$

Form B.1 and B.40 we have

$$L_k^{\tau_S - 1} + \log \frac{p_k}{\mathbb{P}[\nu > \tau_S - 1]} \leq \log \frac{1-\alpha}{\alpha}. \tag{B.3}$$

Let $\eta_\alpha(k) = g^{-1}_{\theta_1,\theta_0}(\log \frac{1-\alpha}{\alpha}; k)$.

$$\sum_{i=k}^{\tau_S - 1} l_i + \log \frac{p_k}{\mathbb{P}[\nu > \tau - 1]} \leq \log \frac{1-\alpha}{\alpha} \leq \sum_{i=k}^{k+\eta_\alpha(k)} I_i. \tag{B.4}$$

For expected delay when $\nu = k$ we have

$$\begin{aligned} \mathbb{E}^{(k)}[(\tau_S - k)^+] &\leq \mathbb{E}^{(k)}[(\tau_S - k - \eta_\alpha(k))^+ + \eta_\alpha(k)] \\ &= \eta_\alpha(k) + \sum_{i=1}^{\infty} \mathbb{P}^{(k)}[\tau_S - k - \eta_\alpha(k) \geq i]. \end{aligned} \tag{B.5}$$

Let $Q_n = \mathbb{P}[\nu > n]$. From B.42, for $i \geq 2$

$$\begin{aligned}
\mathbb{P}^{(k)}[\tau_S - k - \eta_\alpha(k) \geq i] &\leq \mathbb{P}^{(k)}[\sum_{s=k}^{k+\eta_\alpha(k)+i-1} l_s + \log\frac{p_k}{Q_{k+\eta_\alpha(k)+i-1}} \leq \sum_{s=k}^{k+\eta_\alpha(k)} I_s] \\
&\leq \mathbb{P}^{(k)}[\sum_{s=k}^{k+\eta_\alpha(k)+i-1} (l_s - I_s) \\
&\qquad \leq -\log\frac{p_k}{Q_{k+\eta_\alpha(k)+i-1}} - \sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_{\alpha,\epsilon}(k)+i-1} I_s] \\
&\leq \mathbb{P}^{(k)}[\sum_{s=k}^{k+\eta_\alpha(k)+i-1} \lambda(I_s - l_s) \\
&\qquad \geq \lambda\log\frac{p_k}{Q_{k+\eta_\alpha(k)+i-1}} + \lambda\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s] \qquad \text{(B.6)} \\
&\leq \mathbb{P}^{(k)}[\exp(\lambda\sum_{s=k}^{k+\eta_\alpha(k)+i-1} (I_s - l_s)) \\
&\qquad \geq \exp(\lambda\log\frac{p_k}{Q_{k+\eta_\alpha(k)+i-1}} + \lambda\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s)] \qquad \text{(B.7)} \\
&\leq \frac{\mathbb{E}^{(k)}[\exp(\lambda\sum_{s=k}^{k+\eta_\alpha(k)+i-1}(I_s - l_s))]}{\exp(\lambda\log\frac{p_k}{Q_{k+\eta_\alpha(k)+i-1}} + \lambda\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s)} \qquad \text{(B.8)} \\
&\leq \frac{\exp(\lambda^2\sum_{s=k}^{k+\eta_\alpha(k)+i-1} I_i)}{\exp(\lambda\log\frac{p_k}{Q_{k+\eta_\alpha(k)+i-1}} + \lambda\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s)}. \qquad \text{(B.9)}
\end{aligned}$$

To achieve B.44, both sides of the inequality in the probability argument are multiplied by $-\lambda$ for some positive λ. To obtain B.45 an exponential function is applied. The line B.46 is a result of Marcov's inequality. By taking expectation we arrive at B.47.

Let $i_0 = \min\{i \in \mathbb{N} : i \geq -\frac{1}{\xi_2}\log p_k - k - \eta_\alpha(k) + 1\}$. For $i \geq i_0$, by assumption 9.25, we have

$$\log\frac{p_k}{Q_{k+\eta_{\alpha,\epsilon}(k)+i-1}} \geq 0. \qquad \text{(B.10)}$$

Thus, for $i \geq i_0$

$$\begin{aligned}\mathbb{P}^{(k)}[\tau_S - k - \eta_{\alpha,\epsilon}(k) \geq i] &\leq \frac{\exp(\lambda^2 \sum_{s=k}^{k+\eta_\alpha(k)+i-1} I_s)}{\exp(\lambda \sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s)} \\ &\leq \exp(-\frac{(\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s)^2}{4 \sum_{s=k}^{k+\eta_\alpha(k)+i-1} I_s}). \end{aligned} \tag{B.11}$$

To arrive at B.11, let

$$\lambda = \frac{\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s}{2(\sum_{s=k}^{k+\eta_\alpha(k)+i-1} I_s)}. \tag{B.12}$$

From B.43 and B.11, we have

$$\mathbb{E}^{(k)}[(\tau_S - k)^+] \leq \eta_\alpha(k) + i_0 + \sum_{i=i_0+1}^{\infty} \exp(-\frac{(\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s)^2}{4 \sum_{s=k}^{k+\eta_\alpha(k)+i-1} I_s}). \tag{B.13}$$

Next, we calculate the summation on the right hand side of B.53. Starting with the exponent term

$$\begin{aligned}\frac{(\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s)^2}{\sum_{s=k}^{k+\eta_\alpha(k)+i-1} I_s} &= \frac{\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s}{\sum_{s=k}^{k+\eta_\alpha(k)} I_s + \sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s} \sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s \\ &= \frac{1}{\frac{\sum_{s=k}^{k+\eta_\alpha(k)} I_s}{\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s} + 1} \sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s \\ &\geq \frac{1}{\frac{\sum_{s=k}^{k+\eta_\alpha(k)} I_{k+\eta_\alpha(k)}}{\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_{k+\eta_\alpha(k)}} + 1} \sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s \\ &= \frac{1}{\frac{\eta_\alpha(k)}{i-2} + 1} \sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s \\ &= \frac{i-2}{\eta_\alpha(k)+i-2} \sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s \\ &\geq \frac{(i-2)^2}{\eta_\alpha(k)+i-2} I_{k+\eta_\alpha(k)+1}. \end{aligned} \tag{B.14}$$

Substituting the exponent term with B.54

$$\begin{aligned}
&\sum_{s=i_0+1}^{\infty} \exp\left(-\frac{\left(\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s\right)^2}{4\sum_{s=k}^{k+\eta_\alpha(k)+i-1} I_s}\right) \\
\leq &\sum_{i=2}^{\infty} \exp\left(-\frac{(i-2)^2}{4(\eta_\alpha(k)+i-2)} I_{k+\eta_\alpha(k)+1}\right) \\
\leq &\int_{x=0}^{\infty} \exp\left(-\frac{x^2}{4(\eta_\alpha(k)+x)} I_{k+\eta_\alpha(k)+1}\right) dx + 1 \\
\leq &\sqrt{\frac{8\pi\eta_\alpha(k)}{I_{k+\eta_\alpha(k)+1}}} + \frac{8}{I_{k+\eta_\alpha(k)+1}} + 1. \qquad \text{(B.15)}
\end{aligned}$$

Thus, we have

$$E^{(k)}[(\tau_S - k)^+] \leq \eta_\alpha(k)(1+\epsilon(k)) - \frac{1}{\xi_2}\log p_k, \qquad \text{(B.16)}$$

with

$$\epsilon(k) = \sqrt{\frac{8\pi}{\eta_\alpha(k) I_{k+\eta_\alpha(k)+1}}} + \frac{8}{\eta_\alpha(k) I_{k+\eta_\alpha(k)+1}} + \frac{2}{\eta_\alpha(k)}. \qquad \text{(B.17)}$$

Taking expectation over k

$$\begin{aligned}
\mathbb{E}[(\tau_S - \nu)^+] &= \sum_{k=1}^{\infty} p_k \mathbb{E}^{(k)}[(\tau_S - k)^+] \\
&\leq \sum_{k=1}^{\infty} p_k(\eta_\alpha(k)(1+\epsilon(k)) - \frac{1}{\xi_2}\log p_k) \\
&\leq \sum_{k=1}^{\infty} p_k \eta_\alpha(k)(1+\epsilon_0) + \frac{1}{\xi_2} H_p, \qquad \text{(B.18)}
\end{aligned}$$

where H_p is the entropy of the change-point distribution and ϵ_0 goes to zero as α goes to zero. Since $\lambda_{\tau_S} = \frac{\mathbb{P}[\nu \leq \tau_S | X^n]}{\mathbb{P}[\nu > \tau_S | X^n]}$, from $\lambda_{\tau_S} \geq \frac{1-\alpha}{\alpha}$ we have

$$\mathbb{P}[\nu > \tau_S | X^n] \leq \alpha. \qquad \text{(B.19)}$$

Finally, we arrive at the theorem by

$$\mathbb{E}[(\tau_S - \nu) | \tau_S \geq \nu] = \frac{\mathbb{E}[(\tau_S - \nu)^+]}{\mathbb{P}[\tau_S \geq \nu]}$$

$$\leq \frac{1}{1-\alpha}\sum_{k=1}^{\infty} p_k(\eta_\alpha(k)(1+\epsilon_0)) + \frac{1}{\xi_2}H_p$$
$$\leq (1+\epsilon)\sum_{k=1}^{\infty} p_k g^{-1}_{\theta_1,\theta_0}(\log\frac{1-\alpha}{\alpha}; k), \tag{B.20}$$

where ϵ goes to zero as α goes to zero. ■

B.2 Proof of Theorem 26

Let $\eta_{\alpha,\epsilon}(k) = g^{-1}_{\theta_1,\theta_0}((1-\epsilon)\log(1/\alpha); k)$. By Markov's inequality

$$\mathbb{E}^{(k)}[(\tau-k)^+] \geq \eta_{\alpha,\epsilon}(k)\mathbb{P}^{(k)}[\tau - k \geq \eta_{\alpha,\epsilon}(k)]. \tag{B.21}$$

Thus, we have

$$\begin{aligned}
\mathbb{E}[(\tau-\nu)^+] &= \sum_{k=1}^{\infty} p_k \mathbb{E}^{(k)}[(\tau-k)^+] \\
&\geq \sum_{k=1}^{\infty} p_k \eta_{\alpha,\epsilon}(k)\mathbb{P}^{(k)}[\tau - k \geq \eta_{\alpha,\epsilon}(k)] \\
&= \sum_{k=1}^{\infty} p_k \eta_{\alpha,\epsilon}(k)(1 - \mathbb{P}^{(k)}[k \leq \tau < k + \eta_{\alpha,\epsilon}(k)]) \\
&\quad - \sum_{k=1}^{\infty} p_k \eta_{\alpha,\epsilon}(k)\mathbb{P}^{(k)}[\tau < k].
\end{aligned} \tag{B.22}$$

We first show that $\mathbb{P}^{(k)}[k \leq \tau < k + \eta_{\alpha,\epsilon}(k)] \to 0$ as $\alpha \to 0$. Following similar approach as in [76], for some constant $w > 0$,

$$\begin{aligned}
& \mathbb{P}^{(\infty)}[k \leq \tau < k + \eta_{\alpha,\epsilon}(k)] \\
= \; & \mathbb{E}^{(\infty)}[\mathbb{I}(k \leq \tau < k + \eta_{\alpha,\epsilon}(k))] \\
= \; & \mathbb{E}^{(k)}[\mathbb{I}(k \leq \tau < k + \eta_{\alpha,\epsilon}(k))e^{-L_k^\tau}] \\
\geq \; & \mathbb{E}^{(k)}[\mathbb{I}(k \leq \tau < k + \eta_{\alpha,\epsilon}(k), L_k^\tau < w)e^{-L_k^\tau}] \\
\geq \; & e^{-w}\mathbb{P}^{(k)}[k \leq \tau < k + \eta_{\alpha,\epsilon}(k), \max_{k \leq n < k+\eta_{\alpha,\epsilon}(k)} L_k^n < w]
\end{aligned}$$

$$
\begin{aligned}
&\geq e^{-w}(\mathbb{P}^{(k)}[k \leq \tau < k + \eta_{\alpha,\epsilon}(k)] \\
&- \mathbb{P}^{(k)}[\max_{k \leq n < k+\eta_{\alpha,\epsilon}(k)} L_k^n \geq w]). \qquad \text{(B.23)}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\mathbb{P}^{(k)}[k \leq \tau < k + \eta_{\alpha,\epsilon}(k)] \\
\leq\ & e^{w}\mathbb{P}^{(\infty)}[k \leq \tau < k + \eta_{\alpha,\epsilon}(k)] \\
+\ & \mathbb{P}^{(k)}[\max_{k \leq n < k+\eta_{\alpha,\epsilon}(k)} L_k^n \geq w]. \qquad \text{(B.24)}
\end{aligned}
$$

We prove that for $w = (1-\epsilon^2)\log(1/\alpha)$ both terms in the right hand side of inequality B.24 go to zero as α goes to zero.

$$
\begin{aligned}
&\mathbb{P}^{(k)}[\max_{k \leq n < k+\eta_{\alpha,\epsilon}(k)} L_k^n \geq w] \\
&\leq \sum_{n=k}^{k+\eta_{\alpha,\epsilon}(k)-1} \mathbb{P}[\sum_{i=k}^{n} l_i \geq w] \qquad \text{(B.25)} \\
&\leq \sum_{n=k}^{k+\eta_{\alpha,\epsilon}(k)-1} \mathbb{P}[\sum_{i=k}^{n} l_i \geq (1+\epsilon) \sum_{i=k}^{k+\eta_{\alpha,\epsilon}(k)-1} I_i] \\
&\leq \sum_{n=k}^{k+\eta_{\alpha,\epsilon}(k)-1} \mathbb{P}[\sum_{i=k}^{n} (l_i - I_i) \geq \epsilon \sum_{i=k}^{n} I_i + (1+\epsilon) \sum_{i=n+1}^{k+\eta_{\alpha,\epsilon}(k)-1} I_i] \qquad \text{(B.26)} \\
&\leq \sum_{n=k}^{k+\eta_{\alpha,\epsilon}(k)-1} \mathbb{P}[\exp(\lambda \sum_{i=k}^{n} (l_i - I_i)) \\
&\qquad \geq \exp(\lambda\epsilon \sum_{i=k}^{n} I_i + \lambda(1+\epsilon) \sum_{i=n+1}^{k+\eta_{\alpha,\epsilon}(k)-1} I_i)] \\
&\leq \sum_{n=k}^{k+\eta_{\alpha,\epsilon}(k)-1} \frac{\mathbb{E}[\exp(\lambda \sum_{i=k}^{n} (l_i - I_i))]}{\exp(\lambda\epsilon \sum_{i=k}^{n} I_i + \lambda(1+\epsilon) \sum_{i=n+1}^{k+\eta_{\alpha,\epsilon}(k)-1} I_i)} \\
&\leq \sum_{n=k}^{k+\eta_{\alpha,\epsilon}(k)-1} \frac{\exp(\lambda^2 \sum_{i=k}^{n} I_i)}{\exp(\lambda\epsilon \sum_{i=k}^{n} I_i + \lambda(1+\epsilon) \sum_{i=n+1}^{k+\eta_{\alpha,\epsilon}(k)-1} I_i)}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{n=k}^{k+\eta_{\alpha,\epsilon}(k)-1} \exp\left(\frac{-(\epsilon \sum_{i=k}^{n} I_i + (1+\epsilon)\sum_{i=n+1}^{k+\eta_{\alpha,\epsilon}(k)-1} I_i)^2}{4\sum_{i=k}^{n} I_i}\right) \\
&\leq \eta_{\alpha,\epsilon}(k)\exp\left(-\frac{1}{4}\epsilon^2 \sum_{i=k}^{k+\eta_{\alpha,\epsilon}(k)-1} I_i\right) \\
&\leq \eta_{\alpha,\epsilon}(k)\exp\left(-\frac{1}{4}\epsilon^2 \eta_{\alpha,\epsilon}(k) I_k\right). \qquad \text{(B.27)}
\end{aligned}
$$

Inequality B.25 holds by union bound. Following similar steps as proof of Theorem 1 we arrive at B.27. Note that the right hand side of B.27 goes to zero as α goes to zero. From the condition on probability of false alarm

$$
\begin{aligned}
\alpha &\geq \mathbb{P}[\tau \leq \nu] \\
&\geq \mathbb{P}[\tau < n, \nu > n] \\
&= \mathbb{P}[\tau < n | \nu > n]\mathbb{P}[\nu > n] \\
&= \mathbb{P}^{(\infty)}[\tau < n]\mathbb{P}[\nu > n]. \qquad \text{(B.28)}
\end{aligned}
$$

Thus,

$$
\mathbb{P}^{(\infty)}[\tau < k + \eta_{\alpha,\epsilon}(k)] \leq \frac{\alpha}{\mathbb{P}[\nu > k + \eta_{\alpha,\epsilon}(k)]}. \qquad \text{(B.29)}
$$

So, we have

$$
\begin{aligned}
e^{w}\mathbb{P}^{(\infty)}[k \leq \tau < k + \eta_{\alpha,\epsilon}(k)] &\leq \alpha^{\epsilon^2}\frac{1}{\mathbb{P}[\nu > k + \eta_{\alpha,\epsilon}(k)]} \\
&\leq \alpha^{\epsilon^2}\exp(\xi_1(k + \eta_{\alpha,\epsilon}(k))), \qquad \text{(B.30)}
\end{aligned}
$$

which goes to zero as α goes to zero. From B.24, B.27 and B.30, as α goes to zero,

$$
\mathbb{P}_k[k \leq \tau < k + L_\alpha] \to 0. \qquad \text{(B.31)}
$$

Also, note that

$$
\sum_{k=1}^{\infty} p_k \eta_{\alpha,\epsilon}(k)\mathbb{P}[\tau < k] \leq \sum_{k=1}^{\infty} p_k \eta_{\alpha,\epsilon}(1)\mathbb{P}[\tau < k]
$$

$$= \eta_{\alpha,\epsilon}(1)\mathbb{P}[\tau < \nu]$$

$$\leq \alpha\eta_{\alpha,\epsilon}(1). \tag{B.32}$$

Form B.22, B.31, B.32, and the fact that

$$\mathbb{E}[(\tau-\nu)|\tau>\nu] = \frac{\mathbb{E}[(\tau-\nu)^+]}{\mathbb{P}[\tau>\nu]} \geq \mathbb{E}[(\tau-\nu)^+], \tag{B.33}$$

we arrive at the theorem. ■

B.3 Proof of Theorem 27

First we show that the τ_{ML} policy with $\gamma_n = \frac{|\Theta_1|\mathbb{E}[\nu]}{\alpha\mathbb{P}[\nu>n]}$ satisfies the condition on probability of false alarm. Let $\mathcal{S}_k^{(n)}(\phi)$ denote the set of observations that the policy declares a change at $\tau_{ML} = n$, such that $\phi = \arg\sup_{\theta\in\Theta_1} \lambda_n(\theta,\theta_0)$ and $k = \arg\max_{i\leq n} L_i^n(\phi,\theta_0)$.

$$\begin{aligned}
\mathbb{P}^{(\infty)}[\mathcal{S}_k^{(n)}(\phi)] &= \int_{\mathcal{S}_k^{(n)}(\phi)} f(X^{(n)};\theta_0)dX^{(n)} \\
&= \int_{\mathcal{S}_k^{(n)}(\phi)} f(X(1),...,X(k-1);\theta_0)f(X(k),...X(n);\phi) \\
&\qquad \frac{f(X(k),...X(n);\theta_0)}{f(X(k),...X(n);\phi)}dX^{(n)} \\
&\leq \frac{\alpha}{|\Theta_1|\mathbb{E}[\nu]}\int_{\mathcal{S}_k^{(n)}(\phi)} f(X(1),...,X(k-1);\theta_0)f(X(k),...X(n);\phi)dX^{(n)} \qquad \text{(B.34)} \\
&= \frac{\alpha}{|\Theta_1|\mathbb{E}[\nu]}\mathbb{P}^{(k)}_{\theta_0,\phi}[\mathcal{S}_k^{(n)}(\phi)]. \qquad \text{(B.35)}
\end{aligned}$$

The inequality B.34 holds since $L_k^{\tau_{ML}}(\phi,\theta_0) \geq \log\frac{|\Theta_1|\mathbb{E}[\nu]}{\alpha}$. Notation $\mathbb{P}^{(k)}_{\theta_0,\phi}$ denotes the probability operator when the distribution of the first $k-1$ observations is determined by parameter θ_0 and for $t \geq k$ the distribution of observations is

determined by parameter ϕ. Let $\mathcal{U}_k^{(m)}(\phi) = \cup_{n=k}^{m} \mathcal{S}_k^{(n)}(\phi)$.

$$\begin{aligned}
\mathbb{P}^{(\infty)}[\mathcal{U}_k^{(m)}(\phi)] &= \sum_{n=k}^{m} \mathbb{P}^{(\infty)}[\mathcal{S}_k^{(n)}(\phi)] \\
&\leq \frac{\alpha}{|\Theta_1|\mathbb{E}[\nu]} \sum_{n=k}^{m} \mathbb{P}^{(k)}_{\theta_0,\phi}[\mathcal{S}_k^{(n)}(\phi)] \\
&= \frac{\alpha}{|\Theta_1|\mathbb{E}[\nu]} \mathbb{P}^{(k)}_{\theta_0,\phi}[\mathcal{U}_k^{(m)}(\phi)]. \qquad \text{(B.36)}
\end{aligned}$$

Let $\mathcal{U}^{(m)}(\phi) = \cup_{k=1}^{m} \mathcal{U}_k^{(m)}(\phi)$. Thus, we have

$$\begin{aligned}
\mathbb{P}^{(\infty)}[\mathcal{U}^{(m)}(\phi)] &= \sum_{k=1}^{m} \mathbb{P}^{(\infty)}[\mathcal{U}_k^{(m)}(\phi)] \\
&\leq \sum_{k=1}^{m} \frac{\alpha}{|\Theta_1|\mathbb{E}[\nu]} \mathbb{P}^{(k)}_{\theta_0,\phi}[\mathcal{U}_k^{(m)}(\phi)] \\
&\leq \frac{\alpha m}{|\Theta_1|\mathbb{E}[\nu]}. \qquad \text{(B.37)}
\end{aligned}$$

No, we can write

$$\begin{aligned}
\mathbb{P}^{(\infty)}[\tau_{ML} < m] &= \sum_{\phi \in \Theta_1} \mathbb{P}^{(\infty)}[\mathcal{U}^{(m)}(\phi)] \\
&\leq \frac{\alpha m}{\mathbb{E}[\nu]}. \qquad \text{(B.38)}
\end{aligned}$$

From B.38 we can calculate the probability of false alarm

$$\begin{aligned}
\mathbb{P}[\tau_{ML} < \nu] &= \sum_{m=1}^{\infty} p_m \mathbb{P}^{(m)}[\tau_{ML} < m] \\
&= \sum_{m=1}^{\infty} p_m \mathbb{P}^{(\infty)}[\tau_{ML} < m] \\
&\leq \sum_{m=1}^{\infty} p_m \frac{\alpha m}{\mathbb{E}[\nu]} \\
&= \alpha. \qquad \text{(B.39)}
\end{aligned}$$

Following the similar steps as in Theorem 25 and noticing $\lambda_n(\theta_1, \theta_0) \leq \widetilde{\lambda}_n$ we arrive at 9.38. ■

Similar to the proof of Theorem 1

$$\log \lambda_{\tau_S - 1} \leq \log \frac{1-\alpha}{\alpha}. \qquad \text{(B.40)}$$

Form B.1 and B.40 we have

$$L_k^{\tau_S-1}(\theta_1,\theta_0)+\log\frac{p_k}{\mathbb{P}[\nu>\tau_S-1]}\le\log\frac{|\Theta_1|\mathbb{E}[\nu]}{\alpha\mathbb{P}[\nu>\tau_S-1]}. \tag{B.41}$$

Let $\eta_\alpha(k)=g^{-1}_{\theta_1,\theta_0}(\log\frac{|\Theta_1|\mathbb{E}[\nu]}{\alpha};k)$.

$$\begin{aligned}&\sum_{i=k}^{\tau_S-1}l_i+\log\frac{p_k}{\mathbb{P}[\nu>\tau-1]}+\log\mathbb{P}[\nu>\tau_S-1]\\ \le\ &\log\frac{|\Theta_1|\mathbb{E}[\nu]}{\alpha}\le\sum_{i=k}^{k+\eta_\alpha(k)}I_i.\end{aligned} \tag{B.42}$$

For expected delay when $\nu=k$ we have

$$\begin{aligned}\mathbb{E}^{(k)}[(\tau_S-k)^+]&\le\mathbb{E}^{(k)}[(\tau_S-k-\eta_\alpha(k))^+]+\eta_\alpha(k)]\\&=\eta_\alpha(k)+\sum_{i=1}^{\infty}\mathbb{P}^{(k)}[\tau_S-k-\eta_\alpha(k)\ge i].\end{aligned} \tag{B.43}$$

Let $Q_n=\mathbb{P}[\nu>n]$. From B.42, for $i\ge 2$

$$\begin{aligned}\mathbb{P}^{(k)}[\tau_S-k-\eta_\alpha(k)\ge i]&\le\mathbb{P}^{(k)}[\sum_{s=k}^{k+\eta_\alpha(k)+i-1}l_s+\log\frac{p_k}{Q_{k+\eta_\alpha(k)+i-1}}+\log Q\le\sum_{s=k}^{k+\eta_\alpha(k)}I_s]\\&\le\mathbb{P}^{(k)}[\sum_{s=k}^{k+\eta_\alpha(k)+i-1}(l_s-I_s)\\&\qquad\le-\log\frac{p_k}{Q_{k+\eta_\alpha(k)+i-1}}-\lambda\log Q-\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1}I_s]\\&\le\mathbb{P}^{(k)}[\sum_{s=k}^{k+\eta_\alpha(k)+i-1}\lambda(I_s-l_s)\\&\qquad\ge\lambda\log\frac{p_k}{Q_{k+\eta_\alpha(k)+i-1}}+\lambda\log Q+\lambda\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1}I_s] \qquad \text{(B.44)}\\&\le\mathbb{P}^{(k)}[\exp(\lambda\sum_{s=k}^{k+\eta_\alpha(k)+i-1}(I_s-l_s))\\&\qquad\ge\exp(\lambda\log\frac{p_k}{Q_{k+\eta_\alpha(k)+i-1}}+\lambda\log Q+\lambda\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1}I_s)] \qquad \text{(B.45)}\\&\le\frac{\mathbb{E}^{(k)}[\exp(\lambda\sum_{s=k}^{k+\eta_\alpha(k)+i-1}(I_s-l_s))]}{\exp(\lambda\log\frac{p_k}{Q_{k+\eta_\alpha(k)+i-1}}+\lambda\log Q+\lambda\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1}I_s)} \qquad \text{(B.46)}\end{aligned}$$

$$\leq \frac{\exp(\lambda^2 \sum_{s=k}^{k+\eta_\alpha(k)+i-1} I_i)}{\exp(\lambda \log \frac{p_k}{Q_{k+\eta_\alpha(k)+i-1}} + \lambda \log Q + \lambda \sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s)}. \tag{B.47}$$

To achieve B.44, both sides of the inequality in the probability argument are multiplied by $-\lambda$ for some positive λ. To obtain B.45 an exponential function is applied. The line B.46 is a result of Marcov's inequality. By taking expectation we arrive at B.47.

Let $i_0 = \min\{i \in \mathbb{N} : i \geq -\frac{1}{\xi_2} \log p_k - k - \eta_\alpha(k) + 1\}$. For $i \geq i_0$, by assumption 9.25, we have

$$\log \frac{p_k}{Q_{k+\eta_{\alpha,\epsilon}(k)+i-1}} \geq 0. \tag{B.48}$$

Let $i_1 = \min\{i \in \mathbb{N} : i \geq -k - \eta_\alpha(k) + 1 + \max\{\frac{1}{2\xi_2}, \frac{\log 2}{\theta_1}\}\}$. For $i \geq i_1$, by assumption (10), we have

$$I_{k+\eta_\alpha+i-1} + Q_{k+\eta_\alpha+i-1} \geq 0. \tag{B.49}$$

Thus, for $i \geq \max\{i_0, i_1\}$

$$\mathbb{P}^{(k)}[\tau_S - k - \eta_{\alpha,\epsilon}(k) \geq i] \leq \frac{\exp(\lambda^2 \sum_{s=k}^{k+\eta_\alpha(k)+i-1} I_s)}{\exp(\lambda \sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-2} I_s)}$$

Let

$$\lambda = \frac{\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-2} I_s}{2(\sum_{s=k}^{k+\eta_\alpha(k)+i-1} I_s)}. \tag{B.50}$$

$$\mathbb{P}^{(k)}[\tau_S - k - \eta_{\alpha,\epsilon}(k) \geq i] \leq \exp(-\frac{(\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-2} I_s)^2}{4 \sum_{s=k}^{k+\eta_\alpha(k)+i-1} I_s}). \tag{B.51}$$

From B.43 and B.11, we have

$$\begin{aligned}\mathbb{E}^{(k)}[(\tau_S - k)^+] &\leq \eta_\alpha(k) + i_0 + i_1 && \text{(B.52)}\\ &+ \sum_{i=i_0+1}^{\infty} \exp(-\frac{(\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-2} I_s)^2}{4\sum_{s=k}^{k+\eta_\alpha(k)+i-1} I_s}). && \text{(B.53)}\end{aligned}$$

Next, we calculate the summation on the right hand side of B.53. Starting with the exponent term

$$\begin{aligned}&\frac{(\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-2} I_s)^2}{\sum_{s=k}^{k+\eta_\alpha(k)+i-1} I_s}\\ =\ & \frac{\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-2} I_s}{\sum_{s=k}^{k+\eta_\alpha(k)} I_s + \sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-1} I_s} \sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-2} I_s\\ =\ & \frac{1}{\frac{\sum_{s=k}^{k+\eta_\alpha(k)} I_s}{\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-2} I_s} + 1 + \frac{I_{k+\eta_\alpha(k)+i-1}}{\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-2} I_s}} \sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-2} I_s\\ \geq\ & \frac{1}{\frac{\sum_{s=k}^{k+\eta_\alpha(k)} I_{k+\eta_\alpha(k)}}{\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-2} I_{k+\eta_\alpha(k)}} + 1 + \frac{I_{k+\eta_\alpha(k)+i-1}}{I_{k+\eta_\alpha(k)+i-2}}} \sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-2} I_s\\ =\ & \frac{1}{\frac{\eta_\alpha(k)+1}{i-2} + 1 + 4\exp(2\theta_1)} \sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-2} I_s\\ =\ & \frac{i-2}{\eta_\alpha(k) + 1 + (1+4\exp(2\theta_1))(i-2)} \sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-2} I_s\\ \geq\ & \frac{(i-2)^2}{\eta_\alpha(k) + 1 + (1+4\exp(2\theta_1))(i-2)} I_{k+\eta_\alpha(k)+1}. && \text{(B.54)}\end{aligned}$$

Substituting the exponent term with B.54

$$\begin{aligned}&\sum_{s=i_0+1}^{\infty} \exp(-\frac{(\sum_{s=k+\eta_\alpha(k)+1}^{k+\eta_\alpha(k)+i-2} I_s)^2}{4\sum_{s=k}^{k+\eta_\alpha(k)+i-1} I_s})\\ \leq\ & \sum_{i=2}^{\infty} \exp(-\frac{(i-2)^2}{4(\eta_\alpha(k) + 1 + (1+4\exp(2\theta_1))(i-2))} I_{k+\eta_\alpha(k)+1})\\ \leq\ & \int_{x=0}^{\infty} \exp(-\frac{x^2}{4(\eta_\alpha(k) + 1 + (1+4\exp(2\theta_1))x)} I_{k+\eta_\alpha(k)+1}) dx + 1\\ \leq\ & \sqrt{\frac{2\pi(\eta_\alpha(k)+1)}{I_{k+\eta_\alpha(k)+1}}} + \frac{8(1+4\exp(2\theta_1))}{I_{k+\eta_\alpha(k)+1}} + 1. && \text{(B.55)}\end{aligned}$$

Thus, we have

$$E^{(k)}[(\tau_S - k)^+] \leq \eta_\alpha(k)(1+\epsilon(k)) - \frac{1}{\xi_2}\log p_k + \max\{\frac{1}{2\xi_2}, \frac{\log 2}{\theta_1}\}, \tag{B.56}$$

with

$$\epsilon(k) = \sqrt{\frac{8\pi}{\eta_\alpha(k) I_{k+\eta_\alpha(k)+1}}} + \frac{8}{\eta_\alpha(k) I_{k+\eta_\alpha(k)+1}} + \frac{2}{\eta_\alpha(k)}. \tag{B.57}$$

Taking expectation over k

$$\begin{aligned}
\mathbb{E}[(\tau_S - \nu)^+] &= \sum_{k=1}^{\infty} p_k \mathbb{E}^{(k)}[(\tau_S - k)^+] \\
&\leq \sum_{k=1}^{\infty} p_k(\eta_\alpha(k)(1+\epsilon(k)) - \frac{1}{\xi_2}\log p_k + \max\{\frac{1}{2\xi_2}, \frac{\log 2}{\theta_1}\}) \\
&\leq \sum_{k=1}^{\infty} p_k \eta_\alpha(k)(1+\epsilon_0) + \frac{1}{\xi_2} H_p + \max\{\frac{1}{2\xi_2}, \frac{\log 2}{\theta_1}\},
\end{aligned} \tag{B.58}$$

where H_p is the entropy of the change-point distribution and ϵ_0 goes to zero as α goes to zero.

Finally, we arrive at the theorem by

$$\begin{aligned}
\mathbb{E}[(\tau_S - \nu)|\tau_S \geq \nu] &= \frac{\mathbb{E}[(\tau_S - \nu)^+]}{\mathbb{P}[\tau_S \geq \nu]} \\
&\leq \frac{1}{1-\alpha}\sum_{k=1}^{\infty} p_k(\eta_\alpha(k)(1+\epsilon_0)) + \frac{1}{\xi_2} H_p \\
&+ \max\{\frac{1}{2\xi_2}, \frac{\log 2}{\theta_1}\} \\
&\leq (1+\epsilon)\sum_{k=1}^{\infty} p_k g^{-1}_{\theta_1,\theta_0}(\log\frac{1-\alpha}{\alpha}; k),
\end{aligned} \tag{B.59}$$

where ϵ goes to zero as α goes to zero. ■

B.4 Proof of Lemma 9

The proof of Lemma 9 is based on concentration inequalities for Sub-Gaussian distributions.

We prove inequality 10.9 here. The other case, $\mu < \eta$ 10.11, can be proven similarly.

$$\begin{aligned}
&\mathbb{P}\Big[\overline{X}(T) + \sqrt{\frac{2\xi \log \frac{2T^3}{p}}{T}} < \eta\Big] \\
&\quad\leq \mathbb{P}\Big[\sup_s \overline{X}(s) + \sqrt{\frac{2\xi \log \frac{2s^3}{p}}{s}} < \eta\Big] \\
&\quad\leq \sum_{s=1}^{\infty} \mathbb{P}\Big[\overline{X}(s) + \sqrt{\frac{2\xi \log \frac{2s^3}{p}}{s}} < \eta\Big] \\
&\quad\leq \sum_{s=1}^{\infty} \exp(-\log \frac{2s^3}{p}) \qquad \text{(B.60)} \\
&\quad= \sum_{s=1}^{\infty} \frac{p}{2s^3} \\
&\quad\leq p.
\end{aligned}$$

Inequity B.60 is obtained by 10.2.

We next analyze the $\mathbb{E}[T]$ for $\mu > \eta$. Let $s_0 = \min\{s \in \mathbb{N} : \sqrt{\frac{2\log \frac{2s^3}{p}}{s}} \leq \frac{\mu-\eta}{2}, s > 1\}$, for $n \geq s_0$:

$$\begin{aligned}
&\mathbb{P}[T \geq n] \\
&\quad\leq \mathbb{P}\Big[\sup\Big\{s : \overline{X}(s) + \sqrt{\frac{2\xi \log \frac{2s^3}{p}}{s}} > \eta, \text{ and} \\
&\qquad \overline{X}(s) - \sqrt{\frac{2\xi \log \frac{2s^3}{p}}{s}} < \eta\Big\} \geq n\Big] \\
&\quad\leq \mathbb{P}\Big[\sup\Big\{s : \overline{X}(s) - \sqrt{\frac{2\xi \log \frac{2s^3}{p}}{s}} < \eta\Big\} \geq n\Big]
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{s=n}^{\infty} \mathbb{P}\Big[\overline{X}(s) - \sqrt{\frac{2\xi \log \frac{2s^3}{p}}{s}} < \eta\Big] \\
&\leq \sum_{s=n}^{\infty} \mathbb{P}\Big[\overline{X}_s - \mu < -\sqrt{\frac{2 \log \frac{2s^3}{p}}{s}}\Big] \qquad \text{(B.61)} \\
&\leq \sum_{s=n}^{\infty} \exp(-\log \frac{2s^3}{p}) \\
&\leq \sum_{s=n}^{\infty} \frac{p}{2s^3} \\
&\leq \frac{p}{4(n-1)^2}.
\end{aligned}
$$

Notice that B.61 holds because $n \geq s_0$. We can write $\mathbb{E}[T]$ in terms of $\mathbb{P}[T \geq n]$ as

$$
\begin{aligned}
\mathbb{E}[T] &= \sum_{n=0}^{\infty} \mathbb{P}[T \geq n] \qquad \text{(B.62)} \\
&= s_0 + \sum_{n=s_0}^{\infty} \mathbb{P}[T \geq n] \\
&\leq s_0 + \sum_{n=s_0}^{\infty} \frac{p}{4(n-1)^2} \\
&\leq s_0 + 1.
\end{aligned}
$$

For the last inequality notice that s_0 is defined to be bigger than 1. It remains to find s_0. Note that for all $x > 0$ we have $\log x < \sqrt{x}$ so $\log \log x < \log \sqrt{x} = \frac{1}{2} \log x$. For $s = \frac{48}{(\mu-\eta)^2} \log \frac{24 \sqrt[3]{\frac{2}{p}}}{(\mu-\eta)^2}$

$$
\begin{aligned}
\log \frac{2s^3}{p} &= 3 \log \sqrt[3]{\frac{2}{p}} s \\
&= 3 \log \sqrt[3]{\frac{2}{p}} \frac{48}{(\mu-\eta)^2} \log \frac{24 \sqrt[3]{\frac{2}{p}}}{(\mu-\eta)^2} \\
&= 3 \log \sqrt[3]{\frac{2}{p}} \frac{24}{(\mu-\eta)^2} + 3 \log \log \Big(\frac{24 \sqrt[3]{\frac{2}{p}}}{(\mu-\eta)^2} \Big)^2 \\
&\leq 3 \log \sqrt[3]{\frac{2}{p}} \frac{24}{(\mu-\eta)^2} + 3 \log \sqrt[3]{\frac{2}{p}} \frac{24}{(\mu-\eta)^2}
\end{aligned}
$$

$$= 6\log\sqrt[3]{\frac{2}{p}}\frac{24}{(\mu-\eta)^2}$$
$$= 6\frac{(\mu-\eta)^2}{48}s.$$

Thus, for $s = \frac{48}{(\mu-\eta)^2}\log\frac{24\sqrt[3]{\frac{2}{p}}}{(\mu-\eta)^2}$,

$$\sqrt{\frac{2\log\frac{2s^3}{p}}{s}} \leq \frac{\mu-\eta}{2}.$$

So, we have the following upper bound for s_0

$$s_0 \leq \lceil \frac{48}{(\mu-\eta)^2}\log\frac{24\sqrt[3]{\frac{2}{p}}}{(\mu-\eta)^2}\rceil + 1. \tag{B.63}$$

The addition of 1 is because s_0 is defined to be bigger than 1. Thus,

$$\mathbb{E}[T] \leq \frac{48}{(\mu-\eta)^2}\log\frac{24\sqrt[3]{\frac{2}{p}}}{(\mu-\eta)^2} + 2, \tag{B.64}$$

which completes the proof. ■

B.5 Proof of Theorems 28 and 29

An upper bound on the sample complexity of test $\mathcal{A}$ is provided in Lemma 9. Here, we establish an upper bound on the number of times that the test $\mathcal{A}$ is called in CBRW.

In order to analyze the trajectory of the pointer, we consider the last passage time T_l of the pointer from each subtree $\mathcal{T}_l$. We prove an upper bound on $\mathbb{E}[T_l]$ for each l which gives an upper bound on the total number of times the test $\mathcal{A}$

is called. Notice that the total number of times the test $\mathcal{A}$ is called is not bigger than $2\sum_{l=1}^{L}\mathbb{E}[T_l]$.

The pointer initially starts at the root node at distance L from the target and moves as a random walk on the tree. Define the parameters W_t as the steps of the random walk: $W_t = 1$ if the pointer gets one step further from the target at time t, $W_t = -1$ if the pointer gets one step closer to the target, and $W_t = 0$ when the pointer does not move. Clearly, $\sum_{t=1}^{N} W_t = -L$ where N is the stopping time of the random walk. The random walk stops when the policy declares a leaf node as the target. For the mean value of W_t, from Lemma 9, we have

$$\begin{aligned}\mathbb{E}[W_t] &= \mathbb{P}[W_t = 1] - \mathbb{P}[W_t = -1] \\ &\leq 1 - 2(1-p_0)^2 \\ &< 0.\end{aligned}$$

Notice that if the pointer is within the subtree $\mathcal{T}_L$ at step t, we have

$$\sum_{s=1}^{t} W_s > 0. \tag{B.65}$$

Thus, we can write

$$\begin{aligned}\mathbb{P}[T_1 > n] &\leq \mathbb{P}\Big[\sup\{t \geq 1 : \sum_{s=1}^{t} W_s > 0\} > n\Big] &\text{(B.66)}\\ &\leq \sum_{t=n}^{\infty}\mathbb{P}\Big[\sum_{s=1}^{t} W_s > 0\Big] \\ &\leq \sum_{t=n}^{\infty}\exp\Big(-\frac{1}{2}t(1-2(1-p_0)^2)^2\Big) &\text{(B.67)}\\ &= \frac{\exp(-2n(1-2(1-p_0)^2)^2)}{1-\exp(-2(1-2(1-p_0)^2)^2)}.\end{aligned}$$

Inequity B.67 is obtained by Hoeffding inequality for Bernoulli distributions. We can obtain $\mathbb{E}[T_1]$ from $\mathbb{P}[T_1 > n]$ based on the sum of tail probabilities as

$$\mathbb{E}[T_1] = \sum_{n=0}^{\infty}\mathbb{P}[T_1 > n]$$

$$\leq \sum_{n=0}^{\infty} \frac{\exp(-2n(1-2(1-p_0)^2)^2)}{1-\exp(-2(1-2(1-p_0)^2)^2)}$$
$$= \frac{1}{\left(1-\exp(-2(1-2(1-p_0)^2)^2)\right)^2}.$$

Let us define

$$C_{p_0} = \frac{1}{\left(1-\exp(-2(1-2(1-p_0)^2)^2)\right)^2}, \tag{B.68}$$

which is a constant independent of K and ϵ. From the symmetry of binary tree, it can be seen that $\mathbb{E}[T_l] \leq C_{p_0}$ for all l and the expected number of points visited by the pointer is upper bounded by $2LC_{p_0}$. Under the assumption that the informativeness of observations decreases in higher levels we can replace the sample complexity of test $\mathcal{A}$ at the highest level and arrive at the first term in 10.17. The second term in 10.17, is obtained by direct application of Lemma 9 on the sample complexity of test $\mathcal{A}$ at the target node.

It remains to show that CBRW satisfies the reliability constraint. We know that at each visit of a leaf node the probability of declaring a non-target node as the target is lower than $\frac{\epsilon}{2LC_{p_0}}$ by the design of the observation procedure at leaf nodes. Thus from the upper bound on the expected number of points visited by the pointer we have

$$\mathbb{P}[\mathcal{S}_\delta \neq \{(L,1)\}] \leq 2LC_{p_0}\frac{\epsilon}{2LC_{p_0}}$$
$$= \epsilon,$$

which completes the proof of Theorem 28.

An upper bound on the sample complexity of CBRW under the hierarchical setting can be obtained similarly. The trajectory of the pointer can be analyzed

by considering the last passage times T_l of the pointer from subtrees $\mathcal{T}_l$ for $l = l_0+1, ..., L$, as well as the last passage times T'_1 and T'_2 of the pointer from subtrees $\mathcal{T}'_1$ and $\mathcal{T}'_2$ which can be shown to not bigger than $C^H_{p_0}$ following the similar lines as in the proof of Theorem 28 with

$$C^H_{p_0} = \frac{1}{\left(1 - \exp(-2(1-2(1-p_0)^3)^2)\right)^2}. \tag{B.69}$$

The analysis however differs from the analysis of CBRW in that the consecutive calls of test $\mathcal{A}$ on the same node results in increasing the confidence level. We establish an upper bound on the expected total number T_{tot} of observations from a node at a series of consecutive calls of test $\mathcal{A}$ on the node where the confidence level is divided by 2 iteratively at each time test $\mathcal{A}$ is called. Let $T^{(k)}$ be the number of observations taken at k'th consecutive call of test $\mathcal{A}$ on the node. By design of CBRW strategy the value of p in test $\mathcal{A}$ is divided by 2 until the first time k that $\lceil \log_2 \frac{3LC^H_{p_0} p_0}{\epsilon} \rceil$. Thus there are at most $\lceil \log_2 \frac{p_0}{\epsilon} \rceil$ consecutive calls of test $\mathcal{A}$ on one node. On a non-target node:

$$\begin{aligned}
&\mathbb{E}[T_{\text{tot}}] \\
&\quad \leq \sum_{k=1}^{\lceil \log_2 \frac{3LC^H_{p_0} p_0}{\epsilon} \rceil} p_0^{k-1} \mathbb{E}[T^{(k)}] \\
&\quad \leq \sum_{k=1}^{\infty} p_0^{k-1} \left(\frac{48}{(\mu-\eta)^2} \log \frac{24 \sqrt[3]{\frac{2^k}{p_0}}}{(\mu-\eta)^2} + 2 \right) \\
&\quad \leq \sum_{k=1}^{\infty} p_0^{k-1} \left(\frac{48}{(\mu-\eta)^2} \log \frac{24 \sqrt[3]{\frac{2}{p_0}}}{(\mu-\eta)^2} + 2 \right) \\
&\qquad + \sum_{k=1}^{\infty} p_0^{k-1} \frac{48}{(\mu-\eta)^2} \log \sqrt[3]{2^{k-1}}
\end{aligned}$$

$$= \frac{1}{1-p_0}\left(\frac{48}{(\mu-\eta)^2}\log\frac{24\sqrt[3]{\frac{2}{p_0}}}{(\mu-\eta)^2}+2\right) + \frac{p_0}{(1-p_0)^2}\frac{16\log 2}{(\mu-\eta)^2}. \tag{B.70}$$

Upper bound on $\mathbb{E}[T_{\text{tot}}]$ in a conservative upper bound on each single time that the test $\mathcal{A}$ is called.

On the target node:

$$\mathbb{E}[T_{\text{tot}}] \tag{B.71}$$

$$\leq \sum_{k=1}^{\lceil \log_2 \frac{3LC^H_{p_0} p_0}{\epsilon} \rceil} \mathbb{E}[T^{(k)}] \tag{B.72}$$

$$\leq \lceil \log_2 \frac{3LC^H_{p_0} p_0}{\epsilon} \rceil \left(\frac{48}{(\mu-\eta)^2}\log\frac{24\sqrt[3]{\frac{4}{\epsilon}}}{(\mu-\eta)^2}+2\right) \tag{B.73}$$

$$\leq \log_2 \frac{6LC^H_{p_0} p_0}{\epsilon} \left(\frac{48}{(\mu-\eta)^2}\log\frac{24\sqrt[3]{\frac{4}{\epsilon}}}{(\mu-\eta)^2}+2\right). \tag{B.74}$$

From the upper bound on $\mathbb{E}[T_{\text{tot}}]$, the upper bound on the sample complexity of CBRW can be concluded. The satisfaction of the constraint on error probability can be shown similar to Theorem 28. ■.

BIBLIOGRAPHY

[1] T. Lai, H. Robbins, "Asymptotically Efficient Adaptive Allocation Rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4-22, 1985.

[2] R. Agrawal, "Sample Mean Based Index Policies with $O(log n)$ Regret for the Multi-armed Bandit Problem," *Advances in Applied Probability*, vol. 27, pp. 1054-1078, 1995.

[3] P. Auer, N. Cesa-Bianchi, P. Fischer, "Finite-time Analysis of the Multiarmed Bandit Problem," *Machine Learning*, vol. 47, pp. 235-256, 2002.

[4] S. Vakili, K. Liu, Q. Zhao, "Deterministic Sequencing of Exploration and Exploitation for Multi-Armed Bandit Problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 5, pp. 759 - 767, 2013.

[5] P. Auer, R. Ortner, "UCB revisited: Improved Regret Bounds for the Stochastic Multi-armed Bandit Problem," *Periodica Mathematica Hungarica*, vol. 61, no. 1-2, pp. 55-65, September, 2010.

[6] A. Wald, "Sequential Analysis," New York, NY, USA: Wiley, 1947.

[7] A. N. Shiryaev, "The Problem of Quickest Detection of a Violation of Stationary Behavior," *Dokl. Akad. Nauk SSSR,* vol. 138, pp. 1039-1042, 1961.

[8] A. N. Shiryaev, "On optimum Methodes in Quickest Detection Problems," *Theory Prob. Appl.*, vol 8, pp 22-46, 1963.

[9] H. Chernoff, "Sequential design of experiments," *The Annals of Mathematical Statistics*, vol. 30, no. 3, pp. 755770, 1959.

[10] H. Robbins. *Some Aspects of the Sequential Design of Experiments*. Bull. Amer. Math. Soc., vol. 58, no. 5, pp. 527-535, 1952.

[11] S. Bubeck, V. Perchet, P. Rigollet, "Bounded Regret in Stochastic Multi-armed Bandits," arXiv:1302.1611 [math.ST], 2013.

[12] P. Auer, N. Cesa-Bianchi, Y. Freund, R. E. Schapire, "The Nonstochastic Multi-armed Bandit Problem," *SIAM Journal on Computing*, Vol. 32, pp. 48-77, 2003.

[13] A. Garivier, O. Cappe, " The KL-UCB algorithm for bounded stochastic bandits and beyond," *In Proceedings of the 24th Annual Conference on Learning Theory (COLT), JMLR Workshop and Conference Proceedings*, vol. 19, 2011.

[14] K. Liu, Q. Zhao, "Distributed Learning in Multi-Armed Bandit with Multiple Players," *IEEE Transations on Signal Processing*, vol. 58, no. 11, pp. 5667-5681, Nov., 2010.

[15] A. Anandkumar, N. Michael, A.K. Tang, A. Swami, "Distributed Algorithms for Learning and Cognitive Medium Access with Logarithmic Regret," *IEEE JSAC on Advances in Cognitive Radio Networking and Communications*, vol. 29, no. 4, pp. 731-745, Mar., 2011.

[16] Y. Gai and B. Krishnamachari, "Decentralized Online Learning Algorithms for Opportunistic Spectrum Access," *IEEE Global Communications Conference (GLOBECOM 2011)*, Houston, USA, Dec., 2011.

[17] C. Tekin, M. Liu, "Performance and Convergence of Multiuser Online Learning," *Proc. of Internanional Conference on Game Theory for Networks (GAMNETS)*, Apr., 2011,

[18] D. Kalathil, N. Nayyar, R. Jain, "Decentralized learning for multiplayer multi-armed bandits," *Proc. of IEEE Conference on Decision and Control (CDC)*, Dec., 2012.

[19] D. Kalathil, N. Nayyar, R. Jain, "Decentralized learning for multi-player multi-armed bandits," submitted to *IEEE Trans. on Information Theory*, April 2012, available at http://arxiv.org/abs/1206.3582.

[20] S. Bubeck, N. Cesa-Bianchi, "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems," *Foundations and Trends in Machine Learning*, vol. 5, 2012.

[21] K. Liu, Q. Zhao, "Distributed Learning in Multi-Armed Bandit with Multiple Players," *IEEE Trans. on Signal Processing*, vol. 58, no. 11, pp. 5667-5681, November, 2010.

[22] D. Kalathil, N. Nayyar, R. Jain, "Decentralized Learning for Multiplayer Multi-armed Bandits," *IEEE Trans. on Information Theory*, April, 2012.

[23] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial Network Optimization with Unknonw Variables: Multi-Armed Bandits with Linear Rewards", *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1466-1478, October, 2012.

[24] S. Bubeck, N. Cesa-Bianchi, G. Lugosi, "Bandits with Heavy Tail," *IEEE Trans. on Information Theory*, vol. 59, no. 11, Novemeber, 2013.

[25] Y. Deshpande, A. Montanari, "Linear Bandits in High Dimension and Recommendation Systems", *50th Annual Allerton Conference on Communication, Control, and Computing*, 2012.

[26] C. Tekin and M. Liu, "Online Learning Methods in Networking," *Foundations and Trends in Networking*, December, 2014.

[27] M. K. Hanawal, V. Saligrama, M. Valko, R. Munos, "Cheap Bandits", *32th International Conference on Machine Learning*, 2015.

[28] D. Bergemann, J. Valimaki, "Bandit Problems", *Cowles Foundation Discussion Paper*, No. 1551, 2006.

[29] P. Chareka, O. Chareka, S. Kennendy, "Locally Sub-Gaussian Random Variable and the Stong Law of Large Numbers," *Atlantic Electronic Journal of Mathematics*, vol. 1, no. 1, pp. 75-81, 2006.

[30] R. Vershynin, "Introduction to the Non-Asymptotic Analysis of Random Matrices," available at http://arxiv.org/abs/1011.3027v6.

[31] S. Bubeck, N. Cesa-Bianchi, G. Lugosi, "Bandits with heavy tail," *arXiv:1209.1727 [stat.ML]*, September 2012.

[32] Y. Ren, H. Liang "On the best constant in Marcinkiewicz-Zygmund inequality," *Statistics and Probability Letters*, vol. 53, pp. 227-233, June, 2001.

[33] H. Liu, K. Liu, and Q. Zhao, "Learning in A Changing World: Restless Multi-Armed Bandit with Unknown Dynamics," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1902-1916, March 2013.

[34] K. Liu and Q. Zhao, "Adaptive Shortest-Path Routing under Unknown and Stochastically Varying Link States," *Proc. of the 10th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, May, 2012.

[35] Y. Gai and B. Krishnamachari, "Decentralized Online Learning Algorithms for Opportunistic Spectrum Access," Technical Report, March, 2011. Available at http://anrg.usc.edu/www/publications/papers/DMAB2011.pdf.

[36] J. Y. Yu, S. Mannor, "Picewise-stationary Bandirt Problems with Side Observations," *ICML '09 Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1177-1184, 2009.

[37] A. Garivier, E. Moulines, "On Upper Confidence Bound Policies for Non-Stationary Bandit Problems," *Algorithmic Learning Theory*, pp. 174-188, 2011.

[38] P. Tehrani, Q. Zhao, "Stochastic Online Learning under Time-varying Models," *in Proc. of the 46th IEEE Asilomar Conference on Signals, Systems, and Computers*, Nov., 2012.

[39] T. M. Cover, J. A. Thomas, "Elements of Information Theory(Wiley Serie in Telecommunications and Signal Processing),", New York, USA: Wiley, 2006.

[40] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, no. 1 pp. 7791, 1952.

[41] M. C. Steinbach, "Markowitz Revisited: Meanvariance Models in Financial Portfolio Analysis," *SIAM Review* vol. 43, no. 1, pp. 3185, 2001.

[42] A. Sani, A Lazaric, R Munos, "Risk Aversion in Multi-armed Bandits," *Neural Information Processing Systems (NIPS)*, 2012.

[43] R. Yu Ku, "A Study of Risk Aversion: Comparing Contestant Behaviour on Deal or No Deal," Duke University, 2010.

[44] S. Vakili, Q. Zhao, "Mean-Variance and Value at Risk in Multi-Armed Bandit Problems," *53rd Annual Allerton Conference on Communication, Control, and Computing*, 2015.

[45] K. R. French, G. W. Shwert, R. F. Stambaugh, "Expected Stock Returns and Volatility," *Journal of Financial Economics*, no. 19, pp. 3-29, 1987.

[46] E. Even-Dar, M. Kearns, J. Wortman, "Risk-sensitive Online Learning," *17th international conference on Algorithmic Learning Theory (ALT06)*, pp. 199213, 2006.

[47] J. Bradfield, "Introduction to the Economics of Financial Markets", *Oxford University Press*, USA, 2007.

[48] A. Zimin, R. Ibsen-Jensen, K. Chatterjee, "Generalized Risk-Aversion in Stochastic Multi-Armed Bandits," *arXiv:1405.0833[cs]*, 2014.

[49] N. Galichet, M. Sebag, O. Teytaud, "Exploration vs Exploitation vs Safety: Risk-averse Multi-Armed Bandits," *Asian Conference on Machine Learning*, 2013.

[50] O. Maillard, "Robsut Risk-Averse Stochastic Multi-armed Bandits," *Algorithmic Learning Theory*, vol 8139, pp. 218-233.

[51] J. Y. Yu, E. Nikolova, "Sample Complexity of Risk-averse Bandit-arm Selection," *23rd international joint conference on Artificial Intelligence*, pp. 2576-2582, 2013.

[52] L. Tran-Thanh, J. Y. Yu, "Functional Badits," *arXiv:1405.2432 [stat. ML]*, 2014.

[53] J. Audibert, R. Munos, C. Szepesvri, "Exploration-exploitation Trade-off Using Variance Estimates in Multi-armed Bandits," *Theoretical Computer Science*, no. 410, pp. 1876 1902, 2009.

[54] C. Tekin, M. Liu, "Online Learning of Rested and Restless Bandits," IEEE Transaction on Information Theory , vol. 58, issue 8, pp. 5588-5611, Aug., 2012.

[55] W. Dai, Y. Gai, B. Krishnamachari, Q. Zhao, "The Non-Bayesian Restless Multi-armed Bandit: A Case Of Near-Logarithmic Regret," Proc. of Internanional Conference on Acoustics, Speech and Signal Processing (ICASSP) , May, 2011.

[56] C. Tekin, M. Liu, "Approximately Optimal Adaptive Learning in Opportunistic Spectrum Access ," Proc. of International Conference on Computer Communications (INFOCOM) , Mar. 2012, Orlando, Florida USA.

[57] C. Tekin, M. Liu, "Adaptive Learning of Uncontrolled Restless Bandits with Logarithmic Regret," Proc. of Allerton Conference on Communications, Control, and Computing , Sep., 2011.

[58] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial Network Optimization with Unknown Variables: Multi-Armed Bandits with Linear Rewards and Individual Observations," IEEE/ACM Transactions on Networking, vol. 20, no. 5, 2012.

[59] B. Awerbuch, R. Kleinberg, "Online Linear Optimization and Adaptive Routing," Journal of Computer and System Sciences, pp. 97-114, 2008.

[60] Y. Gai, B. Krishnamachari, and M. Liu, "Online learning for combinatorial network optimization with restless markovian rewards." Proc. of the 9th Annual IEEE Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON), 2012.

[61] R. Kleinberg, "Online Decision Problems with Large Strategy Sets," Ph.D. Thesis, MIT, 2005.

[62] K. Liu and Q. Zhao, "Extended UCB1 for Light-Tailed Reward Distributions," available at http://arxiv.org/abs/1112.1768.

[63] P. Chareka, O. Chareka, S. Kennendy, "Locally Sub-Gaussian Random Variable and the Strong Law of Large Numbers," *Atlantic Electronic Journal of Mathematics*, vol. 1, no. 1, pp. 75-81, 2006.

[64] R. Agrawal, "The Continuum-Armed Bandit Problem," *SIAM J. Control and Optimization*, vol. 33, no. 6, pp. 1926-1951, November, 1995.

[65] O. Cappe, A. Garivier, O. Maillard, R. Munos, and G. Stoltz, " Kullback-leibler upper confidence bounds for optimal sequential allocation," *The Annals of Statistics*, 2013.

[66] G. Schwarz, "Asymptotic shapes of Bayes sequential testing regions," *Annals of Mathematics Statistics*, pp. 224-236, 1962.

[67] T. L. Lai, "Nearly optimal sequential tests of composite hypotheses" *the Annal of Statistics*, pp. 856-886, 1988.

[68] T. L. Lai,"Nearly optimal sequential tests of composite hypotheses" *the Annal of Statistics*, pp. 856-886, 1988.

[69] H. Robbins and D. Siegmund,"The expected sample size of some tests of power one"*Annuals Statistics*, pp. 415-436, 1974.

[70] I. V. Pavlov,"Sequential procedure of testing composite hypotheses with applications to the Kiefer-Weiss problem,"*Theory of Probability and its Applications*, vol. 35, no. 2, pp. 280-292, 1990.

[71] A. G. Tartakovsky, "An efficient adaptive sequential procedure for detecting targets," *in Proceeding of IEEE Aerospace Conference*, vol. 4, pp. 1581-1596, 2002.

[72] Y. liu, S. D. Bolstein, "On Optimality of the Sequential Probability Ratio Test for Nonstationary Observations,"*IEEE Transaction on Information Theory*, vol. 38, no. 1, 1992.

[73] A. Tartakovsky, " Asymptotically Optimal Sequntial Tests for Nonhomogeneous Processes," *Sequntial Analysis: Design Methods and Appl*, vol. 17, no. 1, pp. 33-61, 1998.

[74] T. L. Lai, " Asymptotic Optimality of Invariant Sequential Probability Ratio Tests,"*The Annals of Statistics*, vol. 9, no. 2, pp. 318-333, 1981.

[75] W. Shewhart, " Economic Control of Quality of Manufactured Product," New York, NY: Van Nostrand, 1931.

[76] A. G. Tartakovsky, V. V. Veerevalli, "General Asymptotic Bayesinan Theory of Quickest Change Detection," *Theory Prob. Appl.,* vol. 49, no. 3,pp. 458-479, 2005.

[77] K. Cohen and Q. Zhao, "Active hypothesis testing for anomaly detection," *IEEE Trans. on Information Theory*, vol. 61, pp. 1432-1450, 2015.

[78] P. Kundur, J. Paserba, V. Ajjarapu, G. Andersson, A. Bose,C.Canizares, N. Hatziargyriou, D. Hill, A. Stankovic, C. Taylor, T. Van Cutsem, and V. Vittal, "Definition and classification of power system stability IEEE/ CIGRE joint task force on stability terms and definitions," *IEEE Transactions on Power Systems*, vol. 19, no. 3, 2004.

[79] C. Liu, J. S. Thorp, J. Lu, R. J. Thomas, and H. Chiang, "Detection of transiently chaotic swings in power systems using real-time phasor measurements," *IEEE Transactions on Power Systems*, vol. 9, 1994.

[80] S. Dasgupta, M. Paramasiviam, U. Vaidya, V. Ajjarapu, "Real-Time Monitoring of Short-term Voltage Stability Using PMU Data," *IEEE Transactions on Power Systems*, vol. 28, no. 4, 2013.

[81] A. Katok, B. Hasselblatt, " Introduction to the Modern Theory of Dynamical Systems," Cambridge, UK: Cambridge University Press, 1995.

[82] K. Thompson, G. Miller, and R. Wilder, "Wide-area internet traffic patterns and characteristics," *IEEE Network*, vol. 11, pp. 1023, Nov 1997.

[83] P. E. Ayres, H. Sun, H. J. Chao, and W. C. Lau "Alpi: A ddos defense system for high-speed networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 10, pp. 18641876, 2006. 24(10):18641876, 2006.

[84] G. Huang, A. Lall, C. Chuah, and J. X. Uncovering "global icebergs in distributed streams: Results and implications" *Journal of Network and Systems Management*, vol. 19, no. 1, pp. 84110, 2011.

[85] M. Yu, L. Jose, and R. Miao, "Software defined traffic measurement with opensketch.," in *NSDI*, vol. 13, pp. 2942, 2013

[86] R. Dorfman, "The detection of defective members of large populations," in *Annals of Mathematical Statistics*, vol. 14, pp. 436411, 1943.

[87] R. Castro and R. Nowak, "Minimax Bounds for Active Learning," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 2339-2353, 2008.

[88] P. I. Frazier, S. G. Henderson, R. Waeber, "Probabilistic bisection converges almost as quickly as stochastic approximation", available at *arXiv:1612.03964v1 [math.PR]*, 2016.

[89] M. Sobel and P. A. Groll, "Group testing to eliminate efficiently all defectives in a binomial sample," *Bell System Technical Journal*, vol. 38, no. 5, pp. 11791252, 1959.

[90] D.-Z. Du and F. K. Hwang, "Combinatorial group testing and its applications,", World Scientific, vol. 12, 1999

[91] G. Atia and V. Saligrama, "Noisy group testing: An information theoretic perspective," in *47th Annual Allerton Conference on Communication, Control, and Computing,* pp. 355362, IEEE, 2009.

[92] G. K. Atia and V. Saligrama, "Boolean compressed sensing and noisy group testing," *IEEE Transactions on Information Theory*, vol. 58, no. 3, pp. 18801901, 2012.

[93] V. Y. Tan and G. Atia, "Strong impossibility results for noisy group testing," in *ICASSP*, pp. 82578261, 2014.

[94] M. Cheraghchi, A. Hormati, A. Karbasi, and M. Vetterli, "Compressed sensing with probabilistic measurements: A group testing solution," in *47th Annual Allerton Conference on Communication, Control, and Computing*, pp. 3035, IEEE, 2009.

[95] S. Cai, M. Jahangoshahi, M. Bakshi, and S. Jaggi, "Grotesque: noisy group testing (quick and efficient)," in *51st Annual Allerton Conference on Communication, Control, and Computing*, pp. 12341241, IEEE, 2013.

[96] C. L. Chan, S. Jaggi, V. Saligrama, and S. Agnihotri, "Non-adaptive group testing: Explicit bounds and novel algorithms," *IEEE Transactions on Information Theory,* vol. 60, no. 5, pp. 30193035, 2014.

[97] R. Waeber, P. I. Frazier, S. G. Henderson, "Bisection Search With Noisy Responses," in *SIAM Journal on Control and Optimization*, vol. 51, no. 3, pp. 22612279.

[98] M. Ben-Or and A. Hassidim, "The Bayesian learner is optimal for

noisy binary search," in *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science, IEEE*, pp. 221230 ,2008.

[99] R. Castro and R. Nowak, "Active learning and sampling," in *Foundations and Applications of Sensor Management*, Springer, New York, pp. 177200, 2008.

[100] M. Burnashev and K. Zigangirov, "An interval estimation problem for controlled observations," *Problemy Peredachi Informatsii*, vol 10 , pp. 5161, 1974.

[101] G. Cormode, F. Korn, S. Muthukrishnan, and D. Srivastava, "Finding hierarchical heavy hitters in streaming data" *ACM Trans. Knowl. Discov. Data*, vol. 1, no.4, pp 1-48, 2008.

[102] L. Yuan, C. Chuah, and P. Mohapatra, "Progme: Towards programmable network measurement" *In Proceedings of the 2007 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, SIGCOMM, ACM* 2007.

[103] Y. Zhang, S. Singh, S. Sen, N. Duffield, and C. Lund, "Online identification of hierarchical heavy hitters: Algorithms, evaluation, and applications," *In Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement, IMC, ACM,* 2004.

[104] L. Jose, M. Yu, and J. Rexford, "Online measurement of large traffic aggregates on commodity switches" *In Proceedings of the 11th USENIX Conference on Hot Topics in Management of Internet, Cloud, and Enterprise Networks and Services, Hot-ICE, USENIX Association,* 2011.

[105] M. Mitzenmacher, T. Steinke, and J. Thaler, "Hierarchical heavy hitters with the space saving algorithm," *In Proceedings of the Meeting on Algorithm Engineering and Experiments, ALENEX,* 2012.

[106] C. Wang, K. Cohen, Q. Zhao, "Active Hypothesis Testing on A Tree: Anomaly Detection under Hierarchical Observations," to appear in *proceedings of ISIS*, 2017.

[107] S. Chen, T. Lin, I. King, M. Lyu, and W Chen, "Combinatorial pure exploration of multiarmed bandits," in *Advances in Neural Information Processing Systems*, pp. 379387, 2014

[108] A. Locatelli, M. Gutzeit, A. Carpentier, "An optimal algorithm for the Thresholding Bandit Problem," available at: arXiv:1605.08671v1, 2016.